Exploring Implications of Policy Options Concerning Inter-Subject Comparability

ISC Working Paper 6

December 2015

Ofqual/15/5796

Contents

Suggested citation:

Ofqual (2015f) *Exploring Implications of Policy Options Concerning Inter-Subject Comparability: ISC Working Paper 6.* Coventry, the Office of Qualifications and Examinations Regulation.

This report was written by Paul Newton (Research Chair).

1. Introduction

This working paper identifies a range of policy options concerning inter-subject comparability and identifies a variety of issues related to its potential adoption. It focuses specifically upon comparability across the full range of subject areas, leaving open the possibility that slightly different policy considerations might come into play in relation to comparability across certain closely related (that is cognate) subject areas. The policy options focus specifically upon GCSEs and A levels, although similar issues and concerns would arise in relation to other qualification types.

It is worth mentioning at the outset of this working paper that the arguments which it develops are complex. Readers may wish to familiarise themselves with underpinning concepts and underlying debates by reviewing other working papers from this series before starting this one.

Over the years, the bodies that have assumed regulatory responsibilities in relation to GCSE and A level exams in England have all promoted an explicit policy on comparability within subject areas: all necessary action should be taken to achieve comparability within subject areas (over time, between awarding organisations, and so on) through the grade awarding process. Yet, none has promoted an explicit policy on comparability across the full range of subject areas.

The following four statements illustrate the wide range of options that is currently available to us, should we decide that it is appropriate to adopt an explicit policy on inter-subject comparability:

1. No action should be taken to achieve inter-subject comparability through the grade awarding process (which would essentially preserve existing patterns of grade distributions across subject areas).

2. Action should be taken to achieve inter-subject comparability through the grade awarding process (which would lead to different patterns of grade distributions across subject areas than currently exists).

3. No action should be taken to achieve inter-subject comparability through the grade awarding process (which, again, would essentially preserve existing patterns of grade distributions across subject areas), but grades should be scaled subsequently to achieve inter-subject comparability post hoc (and these scaled grades should be reported in addition to the unscaled ones).

4. Action should be taken to achieve a plausible alternative to inter-subject comparability through the grade awarding process (which would lead to different patterns of grade distributions across subject areas than currently exist).

It has already been noted that arguments concerning the strengths and weaknesses of these policy options are complex. One explanation for this complexity is that there is no consensus amongst assessment experts – either nationally or internationally – concerning the definition of inter-subject comparability. In fact, there is no consensus that inter-subject comparability can plausibly be defined at all. The importance of deciding upon a particular definition, as the basis for defending the use of a particular technique for aligning standards, is discussed below. For now, it is important to note that no definition of inter-subject comparability is presumed by this working paper, and that the plausibility of the definitions which are discussed in relation to certain policy options is an important factor in judging the acceptability of those options.

A necessary corollary of this lack of presumption of any particular definition of inter-subject comparability is the lack of presumption that grade standards are currently either aligned or misaligned across GCSE or A level subject areas. In other words, before any judgement can be reached concerning the alignment or misalignment of grading standards, a decision needs to be taken concerning the way in which grading standards are supposed to be aligned, and this is open to debate. Again, this will be discussed in more detail below.

2. Overview of policy options

Before exploring the implications of the policy options in detail, it is useful to begin with a broad overview, to get a sense of how potential justifications might be framed.

1. No action to achieve inter-subject comparability

Comparability between subject areas is far more complicated than comparability within subject areas; not just at a technical level, but also at a conceptual one. In other words, it is very hard to articulate exactly what inter-subject comparability is supposed to mean. In the absence of agreement over how best to define it, let alone agreement over how best to operationalise it, we might decide that we should adopt, as policy, the principle that no action should be taken to achieve it.

In fact, there might be all sorts of justifications for a 'no action' policy. For instance, even if it was possible to define and operationalise inter-subject comparability adequately, if taking action would constitute a threat to intra-subject comparability, and if comparability within subjects was recognised as the highest priority, then it might be better to take no action to achieve comparability across subjects. More generally, if the consequences associated with taking action to achieve inter-subject comparability were judged to be worse, on balance, than the consequences associated with not taking action, then it might be better to take no action to achieve comparability across subjects.

Alternatively, it might even be argued that GCSE and A level exam results do not need to demonstrate inter-subject comparability in order to fulfil their primary purposes. If other organisations (for example, the Department for Education or UCAS) required those results to demonstrate inter-subject comparability to support secondary purposes (for example, accountability or selection), then there might be some justification for recommending that responsibility for scaling results prior to use ought to lie with secondary users.

2. Action to achieve inter-subject comparability

If we considered that it was possible to define and operationalise inter-subject comparability adequately – and if taking action to achieve inter-subject comparability would not unduly threaten higher priority objectives or result in unacceptable consequences – then it would make sense to adopt an explicit policy which required action to achieve inter-subject comparability.

3. Post hoc action to achieve inter-subject comparability

If we considered that it was possible to define and operationalise inter-subject comparability – but concluded that taking action to achieve inter-subject comparability through the grade awarding process might threaten higher priority

objectives or result in unacceptable consequences – then we might consider a more appropriate two-step approach to:

- continue to prioritise intra-subject comparability through the grade awarding process; but
- subsequently scale grades to achieve inter-subject comparability.

In effect, two grades would be provided for each exam: the first engineered to achieve comparability within subjects (over time, across awarding organisations, and so on); the second scaled to achieve comparability across subjects. The implication is that prime and scaled grades ought to be used for different purposes. For instance, scaled A level grades might be dispatched directly to UCAS to be used as the basis for selecting students onto higher education courses.

4. Action to achieve an alternative to inter-subject comparability

Even if we considered that it was either impossible or inappropriate to take action to achieve inter-subject comparability, we might still consider it useful to take some kind of action to change the way in which grades were distributed across subjects. Currently, quite substantial differences exist between subjects in terms of their grade distributions. Although these differences have remained fairly consistent over time – to some extent an inevitable consequence of effective action to achieve comparability over time within subjects – it is not straightforward to provide a convincing explanation for them. For instance, according to the 2014 Joint Council for Qualifications provisional A level statistics, students who took Irish achieved a substantially higher distribution of grades than students who took French, who achieved a substantially higher distribution of grades than students who took Welsh. And, once again, these differentials tend to persist from year to year. Does this imply that, each year, the group of students that takes Irish is substantially ‘better’ than the group which takes French, and the group that takes French is substantially ‘better’ than the group which takes Welsh? If so, then in what sense are they ‘better’? Convincing explanations are hard to come by.

Of course, if it was possible to define and operationalise inter-subject comparability adequately, then this would provide a convincing explanation for differences between grade distributions across subject areas (once the definition had been operationalised). The definition of inter-subject comparability would explain the sense in which students in one subject were ‘better’ than students in the next, and, therefore, why they deserved to be awarded higher grades. If policy option two could be achieved satisfactorily, then no further questions would arise concerning differences in distributions of grades across subjects. If policy option two was not deemed to be viable, though, then a number of alternatives might exist:

- simply carry forward the historical differences in grade distributions with no further justification (for example, policy options one and, to some extent, three); or
- identify an alternative basis for justifying differences in grade distributions across subject areas, plus a mechanism for operationalising that justification; or
- standardise grade distributions across subject areas, such that all subjects are awarded the same distribution of grades.

In theory, alternative bases do exist for justifying differences in grade distributions across subject areas, for example purely on the basis of consequences arising from those differences. However, since none of these bases has yet been worked up into a plausible model, it seems reasonable to focus purely upon standardisation.

Standardisation represents a way of taking action to achieve an alternative to inter-subject comparability, and it would result in different patterns of grade distributions across subject areas than currently exists. It implies that, in the absence of any plausible justification for differences in grade distributions across subject areas, those differences should be eliminated. This approach has been adopted for large-scale tests and exams in various jurisdictions, internationally.

3. Issues that bear upon the policy options

Before reflecting upon the strengths and weaknesses of the various policy options, a number of issues that are relevant to deciding between them will be introduced.

Conceptual issues

To appreciate the viability of the policy options, it is important to recognise that the very idea of inter-subject comparability is highly problematic. The following sections begin by explaining why this is true, before going on to illustrate ways in which some sense might be made of the idea, before considering alternatives to achieving inter-subject comparability.

Comparability within subjects

Comparability is the degree to which results from separate exams embody the same standard. When a high degree of comparability has been achieved between two exams, they are said to be equally difficult, and their results can be used interchangeably. The more similar the constructs (that is characteristics) measured by two exams, the greater the potential to apply the same standard, and the greater the expectation that a high degree of comparability should be achieved.

Perfect comparability would imply that students who were awarded a grade boundary mark on one exam would be exactly the same as students who were awarded the corresponding grade boundary mark on the other exam, in terms of:

- the nature of their attainment (the exam's content); and
- the level of that attainment (the exam's standard).[1]

When a qualification like GCSE geography is provided by two (or more) awarding organisations, it should be possible to achieve a reasonably high degree of inter-subject comparability. However, it would not be possible to achieve perfect comparability. This is because the respective specifications will emphasise different elements of knowledge, skill and understanding in geography – that is, the nature of the geography attainment constructs will differ across the awarding organisations – which means that students who just met the standards set by one awarding

[1] This is 'exactly the same' in an average sense. Even students who are awarded the same mark on the same exam will not be exactly the same as one another in terms of their attainments because they will have different profiles of strengths and weaknesses (that is they will have achieved the same mark total via different routes).

organisation could never be considered exactly the same as students who just met the standards set by another in terms of the nature and level of their attainments.[2]

Comparability across subjects

When making comparisons across qualitatively different qualifications – like GCSE geography and GCSE French – it is obvious that students could never be considered even remotely similar, let alone exactly the same, in terms of the nature and level of their attainments.

Consequently, to be able to achieve inter-subject comparability, the first requirement is to be able to rationalise it somehow. In other words, the first step is to identify some sense in which students who are awarded a grade boundary mark in GCSE geography might be considered the same as, or similar to, students who are awarded the corresponding grade boundary mark in GCSE French. This means that a substantive definition of inter-subject comparability is required – one that is somehow different from the conventional (attainment-based) definition. The second requirement is to be able to operationalise that definition of comparability. In other words, the second step is to identify a technique through which to link standards across subjects, according to the new definition. Historically, debates over inter-subject comparability have been characterised by jumping straight to techniques, with little, if any, attention paid to defining the sense in which students in different subject areas might be considered to be the same or similar. Yet, without first rationalising inter-subject comparability, it is impossible to interpret the 'evidence' that such techniques might provide.

It is important to recognise that there is considerable divergence of opinion amongst assessment experts concerning the legitimacy of the concept of inter-subject comparability, that is the plausibility of any attempt to define it. One leading exam expert once described the concept of inter-subject comparability as "a lunatic idea" (Wood, 1976/1987, p. 42).[3]

Rationalising inter-subject comparability

To achieve, or at least to approximate, inter-subject comparability, we need first to rationalise it. This might well be a 'fuzzy' rationalisation, where the principles and parameters are not defined with absolute precision or certainty. Yet, the rationalisation would need to be sufficiently precise to be able to guide the selection of an appropriate technique and to establish criteria with which to judge whether that

[2] Not even in an average sense. Incidentally, this is also true from year to year within subjects, within awarding organisations, when the content of a specification is substantially revised.
[3] Admittedly, he was writing during a period when the idea of criterion-referencing held far more promise than it ultimately delivered, and this might go some way towards explaining why he felt confident in expressing such an extreme position.

technique was achieving its goal. The underlying logic, methodology and outcomes associated with this approach would all need to be sufficiently plausible to pass the test of public and professional credibility.

As noted above, acting on the basis of an appropriately rationalised technique would require decisions concerning:

- a suitable definition for inter-subject comparability;
- the most appropriate technique for operationalising it.

Various potential definitions and techniques are discussed in the accompanying review of the technical literature (Ofqual, 2015b). Arguably, the most promising pairing so far identified is the combination of:

- defining inter-subject comparability in terms of a linking construct like 'general academic aptitude' or 'general academic application' (that is how effective a student is at learning);
- operationalising this definition using a Rasch-like statistical technique.

In other words:

a) if it is reasonable to believe that some students tend to be more effective at learning than others, such that those differences tend to reveal themselves regardless of the subject studied; and

b) if it is judged to be acceptable to link standards on the basis of this linking construct alone (which would mean 'factoring out' any potential subject-specific effects, such as one subject having more effective teachers than another, or one subject having a higher allocation of curriculum time); then

c) outcomes from Rasch-like statistical techniques would appear to provide a reasonable indication of inter-subject comparability misalignment (with no need for additional untestable assumptions); and, therefore

d) those outcomes would appear to provide a legitimate basis for action to approximate inter-subject comparability.

Premises a) and b) above are certainly debatable. The first premise is clearly not true in an absolute, universal sense. So the debate would focus upon whether it is too gross an oversimplification of reality to form the basis for rational action. Even with the same evidence base (for example, cross-subject grade inter-correlation matrices) individuals might well reach different conclusions on this matter. We would, therefore, need to establish grounds for arbitrating the debate.

Standardisation as an alternative

The alternative that is being considered in relation to policy option four is to standardise grade distributions across subject areas, such that all subjects are awarded the same distribution of grades. This might be viewed as a plausible compromise, or a fall-back position. It recognises the need for some kind of inter-subject grading intervention, whilst acknowledging an inability to achieve inter-subject comparability, per se. The best example of this kind of action is often referred to in England using the term 'norm-referencing', although arguably a better description is provided by the North American expression 'grading on the curve' (that is the bell curve). The basic logic of 'norm-referencing' is to ensure that each cohort of students is differentiated using the same grade distributions. In other words, if 50 per cent of geography students were to be awarded grade C or better, then so too should 50 per cent of French students, biology students, mathematics students, and so on. In other words, the focus is upon standardising grade distributions, rather than on aligning grade standards.

This approach is most informative when there is one national exam per subject per year. In this situation, it can provide very useful comparative information, even in the absence of inter-subject comparability. It allows the qualification user to interpret equivalent grades, from different subjects, to mean that students are of equivalent standing in their respective national cohorts.

In principle, this approach could be adapted to provide the same kind of information, even if there was more than one exam in each subject, each year, which is true in England. First, standards would need to be linked across the set of exams within each subject area according to the conventional definition of comparability; such that students at the same grade boundary marks, across the different exams, would have the same level of attainment in the subject. For each subject, this set would include all exams within the subject area in question; both within and across awarding organisations.[4] Then, the fixed distribution of grades for the subject would be divided out, across the set of exams within that subject area, on the basis of information derived from the conventional linking process.

Technical issues

A variety of technical issues also bear on the choice between policy options. Some of the most important ones are illustrated below.

[4] The details of the approach to linking standards within subject areas would need to be clarified. Most likely, it would involve a combination of methods, including evidence from script judgement exercises and evidence concerning how factors that are known to affect student attainment varied across exam cohorts.

Awarding versus post hoc alignment

Policy option three discusses the possibility of achieving inter-subject comparability post hoc, that is after the grade awarding process has been completed. In effect, this would mean that two grades would be produced for each candidate in each exam – the prime grade and the scaled grade – with the implication that each grade would be fit for a different purpose. As noted above, this scaling might involve the application of a Rasch-like statistical technique. The post hoc scaling would be considered either the responsibility of the awarding organisation (as envisaged by policy option three) or a separate organisation, that is a secondary user of exam results (as discussed in relation to policy option one).

Two specific secondary user bodies spring to mind for GCSE and A level exams:

- the Department for Education (with Ofsted), which collates, distributes and publishes results from qualifications like GCSEs and A levels, in the form of aggregated data, as the basis for evaluating teachers, subject departments, schools, school clusters, and so on;
- UCAS, which collates, distributes and publishes results data from qualifications like A levels, and which provides higher education selectors with individual student results as the basis for making admissions offers and decisions.

If these user bodies were to assume responsibility for the scaling, then no action would need to be taken by awarding organisations to achieve inter-subject comparability, that is the status quo would continue. The awarding organisations would pass exam results directly to the user bodies, which would align standards at the point of use. The goal of point-of-use alignment would be to achieve inter-subject comparability, such that results from exams in different subjects could be interpreted in the same way and used interchangeably.

The critical question from this perspective is how inter-subject comparability should be defined and, consequently, how newly aligned outcomes should be interpreted. In response to a consultation on A level reform, UCAS recently specified the following interpretational requirements:

> HE [higher education] providers need to have confidence that A levels will provide a consistent currency for admissions purposes. This consistency needs to be present in two distinct, but equally important, ways:
>
> - As a consistent indicator of specific subject knowledge (where particular knowledge is a prerequisite for a course).
> - As a consistent general indicator of ability (where HEIs [higher education institutions] are looking for a certain level of attainment rather than subject specific knowledge).

> (UCAS, 2013, p. 2)

The second of these bullets is of particular significance to the inter-subject comparability debate. The idea of a 'general indicator of ability' resonates strongly with the idea of linking standards across subjects on the basis of a linking construct like 'general academic aptitude' or 'general academic application', as discussed earlier. Conceivably, then, UCAS could adopt this definition of comparability, with an appropriate Rasch-like statistical technique, to achieve point-of-use alignment of grade standards across subjects.

In contrast, the Department for Education (with Ofsted) is primarily interested in evaluating teaching effectiveness. Recall that using something like 'general academic aptitude' or 'general academic application' as a linking construct would mean 'factoring out' any potential subject-specific effects, such as one subject having more effective teachers than another. Clearly, from the perspective of evaluating teachers, this would not be an appropriate linking construct. Although a more appropriate linking construct might be identified – such as the 'all causes' definition discussed in *Inter-Subject Comparability: A Review of the Technical Literature* (Ofqual, 2015b) – a technique with the potential to operationalise this kind of definition has not yet been identified. Fortunately, teacher evaluation on the basis of GCSE and A level results data tends nowadays to be carried out within subject areas, on the basis of value-added calculations. That is, teachers are judged in terms of whether they have facilitated the kind of progress in learning that would be expected, for teachers within their subject area.

One question that clearly does arise in relation to performance table data concerns the potential impact of 'easy' subjects on accountability measures which aggregate grades across the profile of subjects taken by students. If it is true that certain subjects are easier to achieve high grades in, and if students from certain schools are disproportionately entered for these subjects, then the aggregated performance profile for those schools would be correspondingly inflated and those schools would appear to be more effective than they actually were. As performance table requirements have been tightened over time, the potential to 'game' the system in this manner has reduced substantially. However, in principle, the Department for Education (with Ofsted) could confront the threat directly through point-of-use alignment to achieve inter-subject comparability of outcomes before they were fed into performance tables. Again, though, the chosen definition of comparability would be critical; as would the match between definition and operational technique. Certain techniques would be just as likely to eliminate relevant differences as irrelevant ones, from the perspective of evaluating teaching effectiveness.

One-off versus ongoing

If it was decided to take some kind of action, it would also need to be decided whether this should involve a one-off alignment, as opposed to an ongoing process.[5] Assuming that awarding organisations carry forward standards effectively from year to year within subject areas, there is some justification for assuming that a one-off alignment across subject areas would subsequently be carried forward over time. So perhaps this is all that would be required?

Having said that, if, for example, attainment standards within a subject area were to rise or fall from year to year, but as a result of teacher factors rather than student factors, then this would threaten inter-subject comparability once again.[6] So, there is certainly a case to be made for an ongoing process rather than a one-off alignment. Of course, the problem with introducing an ongoing process is that whenever such teacher factors operated within a subject area it would establish a direct conflict between the requirement for year-to-year comparability (defined in terms of subject-specific attainment) and the requirement for inter-subject comparability (defined differently).[7]

More generally, it might seem to be incoherent to operate two conflicting definitions of comparability simultaneously – one for year-to-year comparability and one for inter-subject comparability. The rational course of action might, therefore, be to prioritise one at the expense of the other as a matter of principle. Otherwise, whenever the two came into conflict, we would risk the inevitable prioritisation occurring on an unprincipled, random, or capricious basis. This might, for instance, argue in favour of a one-off alignment of inter-subject standards, which was subsequently allowed to deteriorate over time as the year-to-year comparability requirement was prioritised anew.[8] If this was to happen, then new performance descriptions would be required for all subjects to reflect the newly established attainment standards.

The decision to standardise grade distributions across subject areas would not be taken lightly. It would, presumably, be part of a broader reform of the grade awarding system that would imply an ongoing process of re-standardisation, year to year. In

---

[5] Obviously, this question does not arise in relation to point-of-use alignment. This would require an ongoing process of standards maintenance – although 'ongoing recalibration' is perhaps a better term – because the underlying grade standards (reflected in the prime grades) would remain unaligned.
[6] Assuming that they were subject-specific teacher factors, rather than factors which generalised across all subjects.
[7] In theory, recommendations arising from the two definitions would not conflict if an 'all causes' definition of inter-subject comparability was to be adopted. However, as noted earlier, it is not at all clear how this definition could be operationalised.
[8] An alternative course of action could be to define year-to-year comparability differently – specifically, to define it in the same way as for inter-subject comparability – so as to eliminate any potential conflict.

other words, the requirement to achieve year-to-year comparability would also be abandoned. The only remaining requirement would be for comparability within and across awarding organisations for qualifications in the same subject area.

Exceptions

With some of the policy options, it would need to be decided whether they would be rolled out universally, across all subjects in the same way, or whether it would be necessary to treat exceptional cases in different ways. For the standardisation alternative of policy option four, it is hard to see any justification for not standardising universally, so exceptions to this policy will not be considered further.

The question is most relevant to policy options two and three, which recommend action to achieve inter-subject comparability, either during the grade awarding process or post hoc. It has often been said that statistical techniques, like Rasch, can only be used to align subjects that inter-correlate substantially. Although most school curriculum subjects do inter-correlate to a reasonably high degree, some are exceptions. Subjects like art correlate less well with other subjects, and this is also true for more vocationally oriented subjects. Where there is little evidence of inter-correlation, it is hard to provide a rationalisation for alignment based upon the idea of a construct like ‘general academic aptitude’ or ‘general academic application’, which is supposed to reveal itself regardless of the subject studied. Although there would seem to be a good argument for not aligning such subjects with the others, it is not at all clear what action would be appropriate for these exceptional subjects, if the others were to be aligned. This is more than just a problem for the particular ‘outlier’ subjects. It raises serious questions concerning the legitimacy of the alignment policy, per se, when it is not possible to align across the full range of subject areas.

A possible variant of policy option two might be to take action only for subjects that would fall at the extremes of ‘lenience’ and ‘severity’ according to Rasch-like statistical techniques: to bring them closer into alignment with the majority of subjects, without bringing all subjects into exact alignment.

The question of exceptions is also relevant to policy option one, which recommends ‘no action’. There might, for instance, be a small number of clusters of cognate subjects for which it might be considered appropriate to take action to achieve inter-subject comparability; although, this would probably be on the basis of a different definition and technique from that described above in relation to policy option two. As such, this is potentially a situation (as noted at the outset) in which slightly different policy considerations might come into play in relation to comparability across very closely related (that is cognate) subject areas.

Consequential issues

Before commenting on the strengths and weaknesses of each policy option in turn, consequences related to various courses of action will be described at a more general level. We are interested in two different kinds of impacts here – direct and indirect. The primary (direct) reason for wanting to achieve comparability is to ensure that the information provided by exam grades is accurate; more specifically, it is to ensure that equivalent grades convey equivalent information, so that they can be used interchangeably. If they do not convey equivalent information, but are still treated as though they were interchangeable, then this will impact negatively upon the quality of the decisions that the grades are used to make. A secondary (indirect) reason for wanting to achieve comparability is that the perception of non-comparability sends inappropriate signals, which can influence stakeholder behaviour negatively. There is also a somewhat different kind of consequence that needs to be considered: the consequence associated with change.

Information-related impacts

In relation to inter-subject comparability at GCSE and A level, there are at least four different kinds of information users. The principal non-comparability threat for each of them is outlined below (representing the negative impact that would occur if subjects were misaligned):

- If students incorrectly infer that they have the highest aptitude for the subjects in which their grades are the highest, then they may make sub-optimal progression decisions, by pursuing the wrong subjects.
- If selectors incorrectly infer that students with equivalent grade profiles across different subject combinations are the same in terms of ‘general academic aptitude’ or ‘general academic application’, then they may make sub-optimal selection decisions, by appointing the wrong applicants.
- If school managers incorrectly infer that teachers whose students have made the same progress in a subject, from one qualification level to the next, are the same in terms of the ‘value added’ to students, then they may make sub-optimal reward and promotion decisions, by rewarding and promoting the wrong teachers.
- If school evaluators (from Ofsted to parents) incorrectly infer that schools with the same Progress 8 measure are the same in terms of ‘teaching effectiveness’, then they may make sub-optimal evaluation decisions, for example by incorrectly applying to send their children to the wrong schools.

These threats are not necessarily equivalent, in terms of likelihood and significance for their respective information users. For instance, it has already been noted that

results data tend nowadays to be presented to school managers in a way that emphasises the comparison of teachers and subject departments within subjects, rather than comparing across subjects. Similarly, throughout their educational careers, students regularly receive informal feedback on their relative standing in different subject areas. So, a student's final grade in a particular subject is just one piece of evidence amongst many, and it will often arrive only after important progression decisions have been made. Furthermore, in recent years, steps have been taken to minimise the impact that non-comparability would have on performance table outcomes (were non-comparability to exist). So, this threat is no longer as serious as it might previously have been; although the threat still exists.

One use of results that is significantly threatened by non-comparability across subject areas is for selection, either within education or to employment. The most significant threat concerns the use of A level grades for selection to higher education, given the small number of grades with which students will apply, the likelihood that those grades will be from 'similar' subject areas, the fact that the UCAS tariff treats equivalent grades from different subjects interchangeably, and so on. For entry to certain higher education courses, particular A level subject combinations are required, for example three sciences at A level for medicine. When this is true, inter-subject comparability concerns are fairly trivial. However, there are many subjects for which this is not true. Indeed, for entry to certain higher education courses, for example law or psychology, there is typically no requirement to have studied that particular subject at A level and no objection to admitting students with quite different A level specialisms, for example arts versus science backgrounds. Under these circumstances, non-comparability across subjects would present a substantial threat to fair and efficient selection.

Signal-related impacts

Beyond the threat of inaccurate or inefficient decision-making, the perception of non-comparability has the potential to impact negatively in its own right, for instance:

- Students might opt to study 'easier' subjects in favour of 'harder' ones, in an attempt to optimise their trade-off between hard work and high grades.
- Schools might advise students to study 'easier' subjects in favour of 'harder' ones, in an attempt to maximise their performance table outcomes.
- Decision-makers might conclude that they cannot confidently use results for certain purposes, and so base their decisions on alternative indicators which might be less appropriate. Or, they might attempt to 'compensate' for perceived non-comparability, on the basis of hearsay and rumour, by treating students with grades in certain subjects preferentially.

- The public might lose confidence in the value and currency of exam grades, as well as in awarding organisations and the qualifications regulator.

The first of these threats is often considered to be the most serious and has attracted the most attention over time, particularly from subject associations and communities, for example the Association for Language Learning (concerned with the 'flight' from languages at GCSE and A level) and SCORE (Science Community Representing Education) (concerned with the 'flight' from mathematics and science subjects at GCSE and A level). The primary risk is that perceptions of subject difficulty ultimately reduce the supply of students qualified in these subjects to higher-level educational courses and employment. But there are secondary risks, too, such as students becoming increasingly segregated in terms of perceived ability and associated factors.

Having noted the potential negative impacts, it is important also to acknowledge the potential positive impacts associated with perceptions of non-comparability. For instance, it seems likely that students who wish to be perceived as stronger will gravitate towards 'harder' subjects, increasing the concentration of higher-aspiration students in those subjects, and ensuring a more competitive environment for those who are more motivated by competition. Conceivably, then, perceptions of subject difficulty might simultaneously reduce the number of qualified students whilst also increasing their quality. In essentially the same way, lower-aspiration students might gravitate towards 'easier' subjects, ensuring a less competitive environment for those who are less motivated by competition.

Consequences associated with change

If we were to adopt an explicit policy on inter-subject comparability then this would signal change. Even if policy option one (no action) was to be adopted, this would imply that something in the past was wrong, and that the new policy provided a solution.

For any of the options that would change the way in which grades were distributed across subjects, this would raise questions in the minds of those who use exam results and of students and members of the general public. There would be new 'winners' and new 'losers', and these patterns would need to be justified adequately. To the extent that schools and students currently 'game' the system, on the basis of existing folklore surrounding easy and hard subjects, different patterns of gaming would probably arise following a policy change. This might well lead to a substantial period of instability, before stability resumed under the new regime.

4. Strengths and weaknesses of policy options

The following sections provide a succinct summary of many of the most salient strengths and weaknesses related to each policy option. The intention here is to help structure future debate over policy options, rather than to reflect the outcome of a debate that has already been completed. As such, the issues discussed are illustrative, rather than necessarily comprehensive.

1. No action to achieve inter-subject comparability

In a sense, the 'no action' option could be interpreted as the status quo, albeit an implicit status quo. Having said that, the act of formally raising it from implicit to explicit policy might, in itself, have consequences which should be considered.

Strengths

One strong argument in favour of this option is that, despite the 'problem' of inter-subject comparability having been recognised internationally for decades, no country seems yet to have 'solved' it in a manner which has secured widespread public confidence and professional respect. It is not simply that the 'problem' is technically complex, equally it is conceptually complex. Even exam experts disagree over the principles at stake, let alone over the most appropriate techniques to operationalise these principles. It is a highly contested area. If widespread support for any particular course of action could not be secured, then it might be judged appropriate to recommend a policy of no action.

Another argument in favour of no action concerns the threat of unforeseen consequences arising from any of the policy alternatives. The exam system in England is technically very complex, and the stakes associated with exam results are very high. Under these circumstances, unforeseen consequences are likely and, if the anticipated positive impact of policy change is not particularly high, there is a significant risk that the change might cause more harm than good.

In addition to the risk of unforeseen consequences, any action to align/standardise grade standards/distributions across subjects would result in a major change in the currency of exam results for certain subjects, to which students and users would need to acclimatise. It might, for instance, mean many more students with science A levels being admitted to 'high-tariff' universities. It is hard to envisage what the full set of consequences might be, but it is safe to assume that they would be substantial. The consequences might well include new patterns of under- and over-recruitment, with all sorts of resourcing implications, from staffing to accommodation.

It seems likely that any action which involved changing the way in which grades were distributed across subjects would impact upon progression decisions across the full range of subject areas: first for A levels; then for entry to further or higher education.

There might also be related impacts upon progression into different occupations and occupational areas. The precise nature of these changes would be very hard to predict, but it is likely that they would be widespread, affecting many schools, colleges, universities and employers.

Even if those consequences could be justified, as 'appropriate' and 'fair' consequences, they would still require a period of acclimatisation. Any process of 'righting a systemic wrong' would necessarily change the balance of 'winners' and 'losers', and this would require a strong narrative, particularly given the impact on those individuals and groups who would previously have been 'winners' but who would subsequently be 'losers'. However strong the narrative, from a rational point of view, it is likely that it would fail to persuade all stakeholders, particularly those who might have more to lose than to gain from the change.

All of these substantial threats, which would arise in relation to any of the action-related policy options, might be cited in support of a policy of no action. Finally, given the technical complexity of dealing with the challenge of achieving comparability within subjects (between awarding organisations, from year to year, and so on), there is a strong argument for focusing limited technical resources exclusively upon these far more tractable problems, to the exclusion of any attempt to achieve comparability across subjects.

Weaknesses

A major argument against taking no action is that it might be considered tantamount to admitting defeat, particularly if the process of making this policy explicit were to be interpreted (even if incorrectly) as an admission that inter-subject comparability is a genuine concern, but one that cannot be resolved. This might constitute a major threat to public confidence.

This threat might be exacerbated by a continuing inability to provide a convincing justification for differences between grade distributions across subjects, which would carry forward under policy option one. As discussed earlier, if it is judged to be meaningless to conclude that grading standards are misaligned across subject areas, then it must be judged equally meaningless to conclude that grading standards are aligned. This is far from a defence of existing differentials; it is simply an inability to justify any pattern of grade distributions across subjects. In the face of this agnosticism stands the reality that equivalent grades from exams in different subjects are currently treated as though they are interchangeable – particularly for the purpose of selection – and this would presumably continue to occur whether or not any action was to be taken to align grade standards.

An uncomfortable corollary of explicitly adopting the 'no action' policy is that it ought also to be stated explicitly that there is no strong basis for assuming that grades from

different subjects are comparable and, therefore, there is no strong justification for treating them interchangeably. The onus of responsibility would then be upon qualification users to decide for themselves whether or not they were prepared to treat subject grades interchangeably, despite this warning from us.

The 'no action' policy would also do nothing to help ease concerns over the flow of qualified students in the sciences, mathematics, and languages. To be fair, there is no guarantee that aligning grade standards across subjects would necessarily achieve this goal, and, if it did, it would probably have negative impacts on the flow of qualified students to other subjects. So, the significance of this argument against no action – as important as it is – should not be overstated.

Finally, to the extent that it is true that students currently gravitate to particular subjects under the impression that they are 'lenient' and, therefore, easier to achieve high grades in, they would probably continue to do so under the 'no action' policy.

2. Action to achieve inter-subject comparability

If it was decided to align grade standards on the basis of outcomes from a technique based upon Rasch analysis, or something similar, it would also need to be decided in which direction to shift results. This is essentially a political decision, given that there would be pros and cons associated with aligning to the more 'severe' subjects, to the average subjects, or to the more 'lenient' subjects. Engineering all subjects to be as difficult as the most 'severe' ones would mean very large differences to grade distributions in the most 'lenient' subjects, and this would see far fewer students achieving the higher grades. The opposite would happen if all subjects were engineered to be as easy as the most 'lenient' ones.

The most likely decision, given inevitable resistance to radical change, would be to shift all results towards the average. This approach was adopted in the accompanying report *Inter-Subject Comparability of Exam Standards in GCSE and A Level* (Ofqual, 2015c), which modelled the impact of aligning grade standards across subjects according to outcomes from Rasch analyses. Table 1 below illustrates subject-level impacts for some of the more 'lenient' and 'severe' subjects from this study.

As can be seen from table 1, even shifting towards the all-subject average would result in a considerable impact on exam grades towards the extremes. Thus, GCSE German ended up with 11 per cent more students at grade C and above; whilst GCSE English ended up with 18 per cent fewer. At A level, the alignment process resulted in nearly all the further mathematics students (96 per cent) receiving grade C or above.

Table 1: Impact of aligning grade standards across subjects according to outcomes from Rasch analyses on the cumulative per cent of students at grade C (and above).

GCSE English ('lenient')	fall of 18%	from 64% to 46%
GCSE German ('severe')	rise of 11%	from 75% to 86%
A level English ('lenient')	fall of 14%	from 78% to 65%
A level physics ('severe')	rise of 15%	from 74% to 88%
A level further mathematics ('severe')	rise of 6%	from 90% to 96%

Strengths

The strength of this policy option is that it would tackle presumed misalignment of grade standards across subjects directly. For those who considered the underlying logic sufficiently plausible, it would increase confidence in the value and currency of exam grades, and it would enable grades to be treated interchangeably, with a strong justification for doing so. It would, therefore, 'right historical wrongs' and ensure that those students who really deserved high grades were those who actually achieved them, regardless of their subject choices.

To put it simply, if it was possible to achieve widespread consensus over a plausible definition of inter-subject comparability, and if it was possible to identify a technique through which to operationalise that definition adequately, then there might be strong grounds for adopting this policy option.

Weaknesses

The most obvious problem is that this policy option would lead to less differentiation between students in certain subjects, in terms of grades awarded. For some subjects, this would seriously exacerbate an existing problem. In the simulation for A level further mathematics, for instance, the cumulative percentage of students receiving grades A and A* rose from 56 to 67 per cent. This would make it significantly harder for 'high-tariff' universities to select between the best applicants.

Subjects where students were awarded greater proportions of higher grades would be accused of having been 'dumbed down' – this is a media inevitability – and a strong narrative would need to be crafted to confront it. This would be challenging because, of course, grades would have been raised in those subjects with no corresponding rise in attainment levels. It seems quite likely that there would be considerable public dissatisfaction with this apparent grade inflation – however well justified – at least amongst certain communities. This might well include the Russell Group universities, as many of their 'facilitating subjects' would fall into this category.

Just as this policy would increase confidence in the value and currency of exam grades for those who considered the underlying logic sufficiently plausible, it would decrease confidence for those who did not. Conceivably, if views were fairly evenly split, this policy option might have no net impact on public confidence. An outcome like this would be extremely problematic, given the associated costs of the upheaval of implementing the policy option and the significant risk of substantial unanticipated negative consequences.

For some stakeholders, the underlying logic might prove to be especially unconvincing for certain subjects. GCSE Urdu is an oft-used example of a subject for which statistical adjustment might seem unfair. This is a subject in which students typically achieve higher grades than might be expected, given how they attain in the other subjects they study. However, this is often rationalised on the basis that, for many students, this is their family language and it is only natural that they should have superior expertise. This ‘natural advantage’ would be eliminated on the basis of outcomes from a Rasch-like statistical technique.

Exactly the opposite effect would occur for subjects like A levels in critical thinking and general studies. Their relatively low grade distributions are often rationalised on the basis that, for many students, these are low priority subjects to which they exert less effort and commit less time. Again, this ‘natural penalty’ would be eliminated on the basis of outcomes from a Rasch-like technique.

Even a rationalised alignment process would not provide any sort of inter-subject comparability panacea. The alignment would enable users to treat grades interchangeably, but only in the sense given by the linking construct. In the example we have been considering, it would only allow users to infer that students with equivalent grades (from different subjects) were broadly the same in terms of their level of ‘general academic aptitude’ or ‘general academic application’ or some such construct.

Finally, extending the previous point, building inter-subject comparability into grades by linking levels of ‘general academic aptitude’ or ‘general academic application’ would frustrate attempts to build intra-subject comparability into grades by linking levels of ‘subject-specific attainment’. The desire to continue prioritising comparability over time, within subject areas, is one of the main arguments in favour of policy option three over option two.

Big bang versus gradual change

The consequences of adopting policy option two would be substantial, but those consequences might differ somewhat depending on whether the alignment was brought about gradually, over a reasonably long period of time, or was achieved in a single year. There might be arguments for and against either course of action.

The 'big bang' approach might be easiest for awarding organisations to communicate and for stakeholders to understand. The transition from old standard to new would be clear-cut, and the consequences of the transition would be obvious. It might not be easy to adjust to the consequences of such a radical transition, but at least it would be evident to stakeholders exactly what consequences would need to be accommodated.

The 'gradual change' approach might be easiest for students and qualification users to accommodate and acclimatise to. The principle underlying gradual alignment would, presumably, be to ensure that, from year to year, standards changed as little as possible. This might help to ameliorate the impact of radical change on standards within subjects (over time) as well as upon standards across subjects. The aspiration would, presumably, be that qualification users could legitimately treat grades in more or less the same way, from one year to the next, throughout the entire duration of the transition. As the profile of 'winners' and 'losers' gradually changed over time, qualification users would be able to recalibrate gradually the processes and mechanisms that were affected, for example admissions offers.

There are technical issues to consider here, as well as consequential ones. In particular, to accommodate a substantial rise/fall in standards for certain qualifications, their exam papers would need to be made correspondingly harder/easier, to ensure that grade boundaries were not skewed towards the top/bottom of the mark scale (that is to ensure that there continued to be enough marks awarded within each grade band to maintain existing levels of grade reliability). In other words, if A level English grading standards were to be raised, then A level English papers would need to be made harder, and vice versa for A level physics. Changing the difficulty profile of questions within exam papers is not a straightforward exercise (particularly in the absence of 'pre-test' trialling), and the desired transition would be easier to achieve gradually rather than from one year to the next.

3. Post hoc action to achieve inter-subject comparability

Post hoc alignment has been adopted in a handful of education systems around the world, including a number of Australian states and Cyprus, where grades are recalibrated across subjects specifically for the purpose of university admissions (Lamprianou, 2009). The Australian Tertiary Admission Rank (ATAR), for instance, is a numerical measure of a student's overall academic achievement in the Higher School Certificate in relation to that of other students (UAC, 2013). The ATAR is rationalised less in terms of a specific linking construct and more in terms of modelling counter-factual scenarios:

> The ATAR, which aims to provide a fair and equitable method of ranking applicants from all states, is based on the assumption that the age cohorts

> from which the states' Year 12 cohorts are drawn are equally able to undertake tertiary study. That is, if everyone in the age group completed Year 12, it would be fair to consider as admissible to any particular university course the same proportion of each state's students (UAC, 2013, p. 3).

> The scaling algorithm estimates what students' marks would have been if all courses had been studied by all students (UAC, 2013, p. 6).

In practice, the techniques used to model these scenarios share much of their underpinning logic with Rasch-like techniques.

Strengths

The principal advantage of this policy option is that it enables threats from (presumed) inter-subject non-comparability to be addressed without having to add any burden to the grade awarding process. Many of the risks associated with attempting to align standards as part of the process, therefore, disappear. Furthermore, there should be less confusion concerning how (prime) grades ought to be interpreted, as they should only ever be interpreted in the conventional sense, as subject-specific measures of attainment.

More importantly, the scaled grades would be tailored to secondary uses, which would help to maximise their validity in relation to those uses. The fact that different grades (prime versus scaled) would be produced for different purposes should, in theory, eliminate perverse incentives to 'game' the system by choosing to study more 'lenient' subjects.

Weaknesses

A particular concern with policy option three is whether it would reap sufficient benefits to justify the extra expense and confusion that it would cost, in comparison with policy option one or two. The basic logic of policy option three is that it should enable scaled results to be maximally valid for uses which require inter-subject comparability, whilst allowing prime results to remain maximally valid for uses which do not require it. There are problems with both of these suppositions, though.

First, let's assume that only scaled A level results were used as the basis for selection to higher education. If we accept the premise that scaled grades are more interchangeable than prime grades, then this would help to ensure a fairer selection process. On the other hand, we know that applying statistical 'corrections' from Rasch-like techniques would exacerbate problems of differentiation, which have been the subject of much controversy in recent years. Conceivably, then, selectors in

different subject areas or in different institutions might reach different conclusions as to whether the scaled results were actually better for selection purposes than the prime results. If, in response, selectors were encouraged to use either prime or scaled grades (or both), then the system would become far more confusing and potentially open to abuse, either intentionally or unintentionally.

Second, this option is based on the assumption that the current approach embodies features which are especially important to preserve, for primary uses of grades. As we have seen, applying statistical 'corrections' would exacerbate problems of differentiation. On the other hand, the very fact that we already acknowledge problems of grade differentiation raises questions concerning the current system (compared with the standardisation approach, which would ensure optimum grade differentiation across all subjects). Similarly, the increasing reliance upon statistical predictions in recent years also raises questions concerning the current system, this time concerning the feasibility of operationalising an attainment-based definition of comparability in relation to standards over time.

In short, there might be reasons to question whether prime grades currently serve primary purposes well enough to justify insistence that the existing system should necessarily remain untouched, and there might be reasons to question whether scaled grades would serve secondary purposes well enough to justify a two-step approach. These reasons could be marshalled in favour of the more 'economical' policy option two over option three.

Of course, the two-step approach would only make sense if it was possible to achieve some kind of consensus over a plausible definition of inter-subject comparability, and if it was possible to identify a technique through which to operationalise that definition adequately. If either of these pre-requisites could not be achieved, then there would be little, if any, justification for pursuing either policy two or three.

Finally, unless carefully explained and managed, the production of two grades for each exam could end up being extremely confusing for students, users and stakeholders alike. Furthermore, given that there are potentially large numbers of uses to which results might be put, this would raise challenging questions concerning which of the results – prime or scaled – would be most valid for each of these uses. The public understanding threat should not be underestimated.

4. Action to achieve an alternative to inter-subject comparability

This policy option could be operationalised through a 'norm-referencing' approach – at a national level, with outcomes aggregated across awarding organisations – whereby the same grade distributions would be awarded within each subject area. It would necessitate committing substantial resources to linking (that is aligning) grade

standards across specifications within each subject area, both within and across awarding organisations. However, the 'norm-referencing' approach would simultaneously standardise grade distributions from subject-to-subject and year-to-year; so it would kill two birds with one stone, potentially making it no more technically challenging than existing arrangements, and perhaps less challenging.

Strengths

From the perspective of differentiating students equally, in terms of subject grades, this is the perfect option. It would be particularly well-suited to the use of results for selection to higher education; far more so, in this respect, than the 'no action' option.

It would also provide a kind of information that is not actually available under existing arrangements – the relative standing of a student within his or her national cohort for each subject. Under the status quo, according to the 2014 Joint Council for Qualifications provisional A level statistics, students at the 50th percentile in their respective subjects were awarded grade A in Irish, grade B in French and grade C in Welsh. Clearly, under existing arrangements, it is impossible to infer students' relative standing within a subject cohort from A level grades alone. Yet, this might well be a particularly useful source of information for university and employment selectors.

By reporting in terms of deciles or even percentiles (rather than grades) this information could be provided at a fairly fine grain size. This information would be extremely simple for users to understand, an advantage in its own right.[9]

Under the status quo, the fact that grade distributions do differ across subjects invites a substantive interpretation – why else, a user might think, would such differentials exist? Yet, no such substantive interpretation is justified at present. A particularly important strength of the standardisation option is that it would be entirely transparent that grade standards were not aligned in any substantive sense. The problems of treating grades interchangeably across subject areas would also be correspondingly more transparent.

Finally, by straightforwardly standardising grade distributions from subject to subject and year to year, it would bring to an end any speculation that exam standards might be rising or falling. The idea of an exam standard rising or falling would become

[9] 'Norm-referenced' A level results might support inferences of the following kinds: students at the 60th percentile in A level French have achieved a higher level of attainment than 60 per cent of all other A level French students that year; students at or above the 95th percentile in A level law have excelled relative to their cohort (and excelled to the same extent, relatively speaking, as students at or above the 95th percentile in any other subject area); students at the 50th percentile are dead average, relative to their peer group.

irrelevant because grade distributions would simply be standardised. Speculation over rising or falling standards is damaging to public confidence – particularly as it is extremely hard to mount a convincing argument for or against claims like these – so, rendering the debate redundant might well have a positive impact upon confidence in the exam system. Moreover, it would focus attention and resources upon a single comparability consideration, rather than spreading them across a range of considerations. It would explicitly prioritise comparability between the same qualifications awarded by different awarding organisations.

Weaknesses

The obvious weakness of this approach is that grade standards would not be aligned across subject areas, or from year to year, in any substantive sense (although they would, of course, be aligned within subject areas, for example between awarding organisations, on the basis of the traditional attainment-based definition of comparability.) The most significant consequence here is that they would not be aligned from year to year. To be fair, the increased prominence of statistical predictions within current grade awarding procedures is an explicit response to a widespread lack of confidence in the ability of awarding organisations to maintain exam standards over time. Yet, the official adoption of a 'norm-referencing' methodology would rule out this objective as a matter of policy.[10]

As for the 'no action' policy option, applying a fairly arbitrary standardisation might be considered tantamount to admitting defeat. This, in turn, might constitute a major threat to public confidence, as discussed above.

Similarly, to the extent that it is true that students currently gravitate to particular subjects under the impression that they are 'lenient' and, therefore, easier to achieve high grades in, they would be likely to do so under a standardisation policy, although a different folklore would no doubt evolve with different kinds of gaming principles.

No explicit policy

Finally, it is important to recognise that continuing not to adopt an explicit policy on inter-subject comparability is still a possibility. This might be justified by arguing that the exam system has managed to function adequately for many decades in the absence of an explicit policy. We might, for instance, conclude that the challenges associated with defining and operationalising inter-subject comparability are insurmountable. If so, then an explicit policy on inter-subject comparability might be

[10] In fact, in large subject areas, with fairly stable cohorts, it is quite likely that students at the 60th percentile in one year would have the same level of attainment as students at the 60th percentile in the next. So, in practice, for many if not most subjects, students at equivalent grades from year to year would be likely to have comparable levels of attainment. Such inferences might, therefore, still be drawn, even though the system would not specifically be designed to support them.

judged to be inappropriate. Most of the strengths and weaknesses associated with this possibility are the same as for policy option one.

Strengths

The ‘no policy’ option reflects the existing status quo. In the absence of a sufficiently convincing case for change, it might well be considered the most appropriate course of action. It has the advantage of being the devil we know.

Weaknesses

The longer that inter-subject comparability remains a topic for public debate, the less sustainable it possibly becomes for the qualifications regulator not to adopt an explicit policy on it. Particularly given the amount of work that we have recently undertaken in this area, the decision to continue with no explicit policy might lack credibility.

5. References

Lamprianou, I. (2009) *Comparability of examination standards between subjects: an international perspective*. Oxford, Oxford Review of Education, 35 (2), pp. 205–226.

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

Ofqual (2015c) *Inter-Subject Comparability of Exam Standards in GCSE and A Level: ISC Working Paper 3.* Coventry, the Office of Qualifications and Examinations Regulation.

Universities Admissions Centre (2013) *Report on the Scaling of the 2012 NSW Higher School Certificate*. Sydney, Universities Admissions Centre.

UCAS (2013) *UCAS response to Ofqual A level Reform consultation*. Cheltenham, UCAS.

Wood, R. (1976/1987) *Your chemistry equals my French*. Letter to the Times Educational Supplement on 30th July. Reprinted in Wood, R. (1987) *Measurement and Assessment in Education and Psychology: Collected Papers 1967–87* (pp. 40–4). Lewes, the Falmer Press.

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