RESEARCH AND ANALYSIS

Maintaining Standards

During normal times and when qualifications are reformed

ofqual

Authors

This report was written by Paul E. Newton, from Ofqual's Strategy, Risk, and Research Directorate.

Contents

1 Introduction

It is far from simple to maintain exam standards during **normal times**. Even when exam boards correctly follow appropriate procedures, errors still sometimes occur. It is even harder to maintain them during periods of **qualification reform**, when curriculum, syllabus, and assessment arrangements change.

Even the ideas that we use to explain what we are trying to achieve when we attempt to maintain standards can be confusing – Attainment-Referencing, Comparable Outcomes, the Sawtooth Effect. More worryingly still, these ideas often appear to become confused, when they are used to scaffold public debate on examining practices, via social media or traditional media outlets.

This report is an attempt to bring all of these ideas together – as clearly and cogently as possible – in an attempt to explain how they relate to each other, and to provide a comprehensive overview of the maintenance of standards territory.

During normal times, we tend to view the maintenance of standards primarily through the lens of **meaning**; that is, in terms of how grades need to be **interpreted**. We strive to ensure that equivalent grades, across successive versions of the same subject exam, can be **interpreted** in the same way; that is, in terms of attainment. We say that exam standards have been maintained when equivalent grade boundary marks across adjacent exams correspond to equivalent levels of attainment. We call this principle Attainment-Referencing.

Conversely, across periods of qualification reform, we tend to view the maintenance of standards primarily through the lens of **consequence**; that is, in terms of how grades will be **used**. We strive to ensure that equivalent grades, across successive versions of the same subject exam – pre-reform versus post-reform – can be used in the same way and used **fairly**.

We switch our focus from meaning to consequence – from the last exam pre-reform to the first exam post-reform – because we expect the quality of candidates' exam performances to drop significantly across this period of transition. It is not the fault of the first cohort post-reform that their performances (and, indeed, their attainments) are lower. This is essentially a consequence of their teachers not yet being up to speed with teaching the new content elements, or with preparing for the new assessment structure/formats. So, we compensate for this effect – the Sawtooth Effect, which represents a sudden dip in performance followed by a gradual rise back up again over time – by applying the Comparable Outcomes principle. We do so in order to be fair to candidates.

The following sections explain these ideas in more depth. They explain the logic of what we are trying to achieve, when we attempt to maintain exam standards, as well as the methods that we use to maintain them. This report is intended as an

introductory overview. A more detailed and nuanced account of grade awarding during normal times can be found in Taylor & Opposs (2018); while a more detailed and nuanced account of Attainment-Referencing, the Sawtooth Effect, and the Comparable Outcomes principle can be found in Newton (2020).

Finally, it is worth mentioning that COVID-19 times are even less normal than periods of qualification reform. However, very similar considerations arise, especially the need to be fair to candidates, and similar methods can be employed as well. However, because the effects of COVID-19 will be far more severe for some learners than for others, applying the Comparable Outcomes principle can only represent a partial solution to the problem of learning loss attributable to COVID-19.

2 Maintaining Standards

If we were to introduce a brand new qualification for first examination in 2021 – let's call it a Z level in Pedagogy – there would be all sorts of design decisions to take. We would need to characterise the relevant body of knowledge and skills for teaching and learning – which typically involves writing a syllabus – and we would need to decide how to assess the degree to which each learner had mastered this domain of learning.

Were we to decide that it ought to be assessed via an external exam, we would need to decide exactly how that exam ought to be constructed. This would involve preparing a **blueprint** to specify:

- the number of marks that would be available (in total)
- the number of tasks/questions that would be set
- the formats that would be used to configure those tasks/questions
- how the body of knowledge and skills would be sampled, including how to weight each element
- how those weights would be distributed across formats
- and so on.

The exam blueprint helps to ensure that the same domain of learning is assessed in the same way each year, which is the foundation for ensuring that exam standards can be maintained over time.

Exams typically operate via a two-step process which involves:

1. **ranking** candidates in terms of their overall level of attainment in the domain of learning, based upon the total number of marks that they achieve in the exam; and then
2. **classifying** candidates in terms of whether their level of attainment is high enough to be awarded a particular grade.

This second step involves locating grade boundaries on the exam mark scale. These **grade boundary marks** divide the scale into grade bands – A*, A, B, and so on – and each candidate is awarded a grade depending on which band their mark total

falls into. The grade boundary mark is therefore the lowest mark associated with each grade.

2.1 Setting exam standards

Grade boundary marks are not decided until after an exam has been sat, once most (if not all) of the exam scripts have been marked. They are the outcome of a process that is known as **grade awarding**; because it is this process that enables candidates' grades to be awarded, based upon the mark that they achieve in the exam.

Exam standards are not officially **set** until the very first time that grade boundary marks are decided for a new qualification. As such, the very first batch of grade boundary decisions establishes the standards that will need to be maintained from then on.

The standard for each exam grade can be defined as the **level of attainment** that is associated with its grade boundary mark. We work on the assumption that each mark on the mark scale (from zero to the total mark for the exam) can be described in terms of a specific level of attainment. This is the level of attainment that it takes to score that many marks on the exam.

The standard for each exam grade can be defined as the level of attainment that is associated with its grade boundary mark.

For the purpose of setting standards and awarding grades, we assume that all candidates who are awarded the same number of marks share the same level of attainment.[1] Hence, the level of attainment that is common to candidates at the grade boundary mark constitutes the exam standard for that grade.

2.2 Maintaining exam standards

As noted above, each new form of an exam is built according to the same blueprint. The 2022 Pedagogy Z level exam would therefore look exactly the same as the 2021 Pedagogy Z level exam, in terms of its structure and formats. Only the specific content of each question would change. Bearing this in mind, it is tempting to

[1] Measurement error means that, in practice, this will not actually be true. So, it is more accurate to say that the grade standard corresponds to the average level of attainment of candidates at the grade boundary mark.

assume that, to apply the same exam standards from one year to the next, we would simply apply the same grade boundaries.

Unfortunately, it is not quite as simple as this. Even when questions are written to assess exactly the same elements of knowledge or skill (from one year to the next) each new question may turn out to be slightly easier or slightly harder than its counterpart from the previous year. Unless these subtle differences in difficulty happen to cancel out, across questions, the overall exam may also turn out to be somewhat easier or harder from one year to the next.

To express this slightly differently, if the new version of an exam turns out to be harder than the old version, then a candidate who scored X marks on the old version would be likely to score less-than-X marks on the new one. More generally, candidates with equivalent levels of attainment, from one year to the next, would be likely to achieve different mark totals, i.e. they would typically achieve a lower mark on the new exam.

When exams do prove to be easier or harder, from one year to the next, we compensate for this by locating grade boundaries at different marks. If, for instance, the new exam turned out to be two marks harder, on average, then we would locate grade boundaries two marks lower to compensate for this. This process enables us to apply exactly the same exam standard, from one year to the next, even when one exam happens to be somewhat easier or harder than the other. We refer to this process as **maintaining** exam standards.

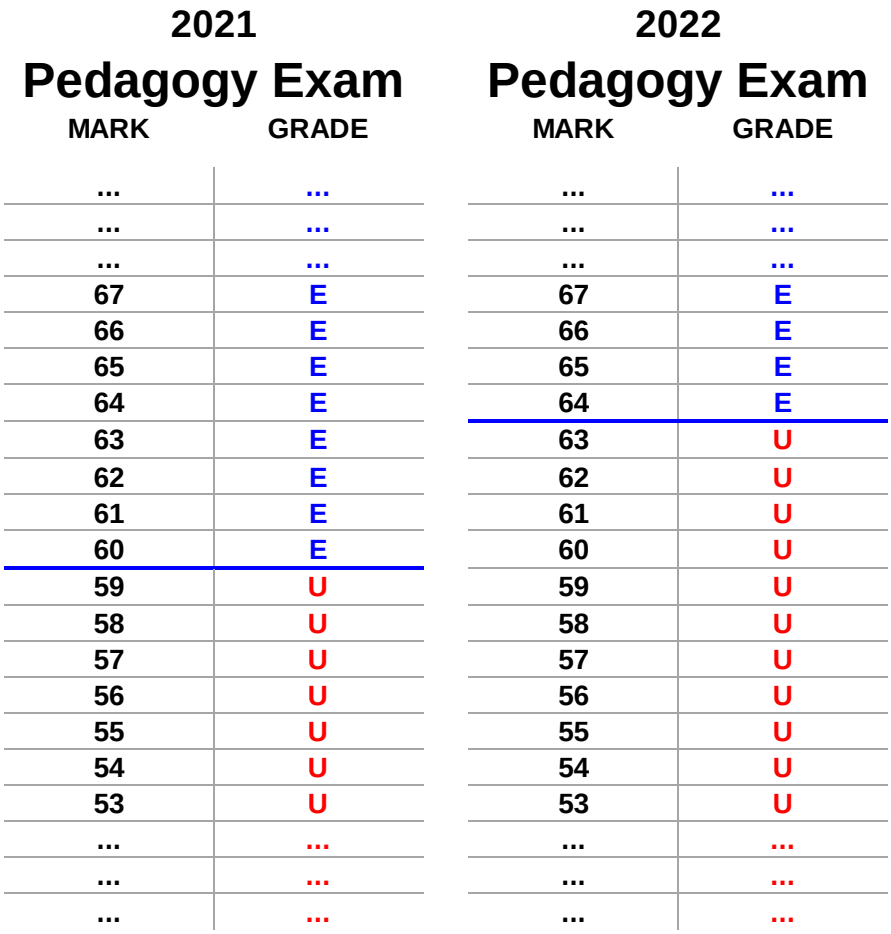

Figure 1. Illustration of an E grade boundary (2020 vs. 2021)

Figure 1 illustrates this process graphically. Here, it was decided that the 2022 exam was somewhat easier for candidates in the lower attainment range than the 2021 exam had been – more precisely, 4 marks easier – so the 2022 grade E boundary was raised by 4 marks to compensate for this. Thus, the grade E exam standard was maintained, i.e. **carried forward**, from 2021 to 2022.

Exam standards are 'carried forward' by locating grade boundaries at marks that correspond to equivalent levels of attainment.

Each subject's exam standards are carried forward from one year to the next by locating grade boundaries at marks that correspond to equivalent levels of

attainment. Because exam standards are referenced to levels of attainment, we call this principle **Attainment-Referencing**.[2]

2.3 Methods for maintaining standards

A good way to determine whether the new form of an exam was more or less difficult than the old form (or of equivalent difficulty) would be to set up an experiment. For instance, we might administer both forms to a single group of candidates, under the same conditions that the exam would normally be sat, and see how they perform. If candidates of all levels of attainment tended, on average, to perform 3 marks better on the new form, then we could conclude that the new form was 3 marks easier, and locate each of its grade boundaries 3 marks lower to compensate for this. When effectively controlled and administered, this can be a very robust approach to maintaining standards. It can also enable us to determine grade boundary marks, on the basis of the experimental trial, well in advance of the exam being sat live.

In England, we tend not to adopt this experimental method, for high stakes exams like GCSEs and A levels, for a variety of reasons including costs. The main reason, though, is that each exam paper needs to be kept strictly confidential until the day of the exam. Running trials before an exam goes live risks security breaches, which would threaten its validity.

Instead, the method that we use in England is not implemented until after an exam has been sat, once most (if not all) of the exam scripts have been marked. It revolves around the work of a **grade awarding committee**, which represents the most senior examiners for the exam in question. They will tend to be teachers, retired teachers, university lecturers, or suchlike, with many years of experience of examining the subject. As experts in the domain, as well as in the exam, they will have acquired a strong sense of what the grade standards 'look like' in terms of the quality of work that they would expect to see at each grade boundary.

The work of the grade awarding committee revolves around a **script scrutiny** exercise. This involves comparing work within scripts from the current exam – work from multiple candidates at a range of marks – with work within reference scripts from previous exams at grade boundary marks.

For example, if the grade C boundary mark in previous years tended to hover at around 120 marks, the committee might be given work from the present exam at a range of marks, say, from 118 to 122. For each of these mark points, there might be work from 5 candidates available for scrutiny. By looking at a large sample of scripts, from across this 5 mark range, they would be aiming to identify the mark at which the

[2] This is sometimes referred to as 'Criterion-Referencing' although that is not actually a very helpful characterisation. The term 'Weak Criterion-Referencing' is perhaps more accurate, although the implication of the prefix is simply that it is 'not really' Criterion-Referencing, in any strong sense.

intrinsic qualities of performances on the current exam most closely resembled the intrinsic qualities of performances *at grade boundary marks* on previous exams. Ultimately, the mark of closest resemblance, on the current exam, would be chosen as the recommended boundary mark, for the grade in question. By the same process, other grade boundary marks would be chosen.[3]

2.4 Similar Cohort Adage

If the judgement of expert examiners was sufficiently precise, nothing more than this method would be required in order to locate grade boundaries. Unfortunately, examiner judgement is not that precise; and it can be rendered even less accurate by judgemental biases, and group dynamics. Consequently, this approach is not as robust as the experimental one described above, and recommendations based upon expert judgement need to be supplemented by additional sources of evidence concerning the likely location of grade boundaries, to enable some kind of triangulation.

For the best part of a century, exam boards in England have relied upon **examiner judgement** of performance evidence, from completed exam scripts, as a key source of evidence for maintaining exam standards. However, they have also always relied upon **statistical expectations** of cohort attainment – based upon assumptions concerning the stability of each subject cohort – in order to supplement examiner judgement. The logic of relying upon statistical expectations goes like this:

> The candidates are not like a fruit crop, which may suffer a blight and produce poor results in any one year; in normal times variations in standard are small, and we should err very little if we kept the percentage of passes in the important subjects fairly constant from year to year.
>
> (Crofts & Caradog Jones, 1928, p.45)

It was well-known, even back in the 1920s, that expert examiners can sometimes get it wrong – badly wrong – when it comes to maintaining exam standards. It is part of the job of those who write an exam to make it as similar in difficulty to the previous version as they can possibly manage. Sometimes, though, for reasons that may never fully come to light, the new form of an exam can turn out to be very different in difficulty from the old form. When this happens, it can be hard for the grade awarding committee, which includes those who were responsible for writing the paper, to appreciate just how easy or hard it turned out to be. They may be prepared to allow a certain amount of compensation, in their grade boundary recommendations, but

[3] Because this process is very time consuming, it tends only to be undertaken for certain grade boundaries, e.g. grade A and grade E. The remaining grade boundaries would be identified by a process of interpolation (equal division of marks between these reference points) or extrapolation.

often not enough. When this happens, pass rates can rise or fall radically, from one year to the next, with little justification.

James Crofts popularised what came to be described as 'the curve' method, as an antidote to the fallibility of examiner judgement.[4] The logic of the curve is simply that the likelihood of a large cohort changing radically from one year to the next, in terms of its average level of attainment, is far less than the likelihood of the new exam paper being substantially easier or harder than the old one. Under such circumstances, we would 'err less' were we to keep the percentages of passes consistent from one year to the next, even if it meant dropping or raising grade boundaries substantially. This way of thinking is encapsulated within a rule-of-thumb, which has been referred to as the **Similar Cohort Adage**, which states that: if the cohort hasn't changed much, then don't expect the pass rate to change much either.

In short, if the cohort for the new exam is fairly large, and demographically similar to the cohort for the old exam, then the chances are that its overall level of attainment will be similar too. After all, we would not expect a cohort of learners to become *substantially* better at learning from just one year to the next; nor would we expect their teachers, en masse, to become *substantially* better at teaching (see Coe, 2013). According to this logic, if grade boundaries were located in order to return the same percentages of candidates at each grade, then we would probably *not err too much*. This process has come to be known as establishing '**statistically expected boundaries**' (SEBs).

SEBs can be used (for example) to determine mid-points for the range of scripts that will be scrutinised by a grade awarding committee; with an expectation that their grade boundary recommendations should not depart substantially from these anchor points, without good justification.

If the cohort hasn't changed much from one year to the next, then don't expect the pass rate to change much either.

Grade awarding in England has always relied upon a judicious balance of expert judgement of performance evidence and statistical expectations of cohort attainment. When examiners recommend grade boundaries that would represent a substantial departure from statistical expectations – which they are at liberty to do – they are

[4] Crofts was Secretary to the Joint Matriculation Board (JMB) of the Northern Universities, from 1919 to 1941.

required also to provide some kind of justification for this. This would include features of performances in scripts, which they considered to be particularly compelling. Ideally, it would also include additional, independent evidence.

In previous decades, independent evidence of the demography of the cohort having changed substantially would have been considered highly relevant (e.g. more girls in the new cohort, or more independent schools). This might well have persuaded the exam board accountable officer – who is ultimately accountable for all grade boundary decisions within the board – to accept recommendations from the awarding committee, where these diverged substantially from previous years.

2.5 Prediction matrices

Nowadays, the judgement of expert examiners is supplemented by a variety of sources of evidence: technical data; investigation outcomes; descriptions of performance standards; and so on. Yet, statistical expectations of cohort attainment still carry considerable weight. More importantly, these expectations are now far more sophisticated than in previous decades, being based upon outcomes from **prediction matrices**.

Prediction matrices enable exam boards to generate statistical expectations that take into account how a subject cohort might have changed from one year to the next (rather than being based on an assumption that the cohorts are very similar). They do so by controlling for the single best predictor of educational attainment – *prior* educational attainment.

Without going into the technical detail of how prediction matrices work, they basically take the percentage of the subject cohort that was awarded each grade on last year's exam, and carry it forward to this year's exam, after having controlled for the calibre of this year's cohort. These prior-attainment-adjusted percentages can then be used to determine SEBs for this year's exam, or 'statistically recommended boundaries' (SRBs) as they are often known nowadays (e.g. CERP, undated).

GCSE predictions are based upon average key stage 2 test results, while A level predictions are based upon average GCSE results; meaning that the calibre of each cohort is judged in terms of an average of candidates' average test or GCSE results. If the mean (average) GCSE score for this year's cohort is lower than the mean (average) GCSE score for last year's cohort, then – all other things being equal – we would expect this year's cohort to perform less well at A level. We use prediction matrices to tell us how much less well.[5]

[5] Because we are using these average test/exam results to measure the calibre of a cohort (not its overall level of attainment per se) we re-standardise them before using them as inputs to prediction matrices. This ensures that we do not inadvertently build grade inflation into the modelling process, if the prior attainment results happened to be affected by Sawtooth-like effects.

Of course, it is quite possible that all other things do not remain equal. For instance, this year's cohort might, on average, have been taught a little better. Or they might, on average, have put a little more effort into their studies. If so, then we would expect them to achieve somewhat better results than our prediction matrices would suggest. And we would hope that our expert judges would pick this up, during their script scrutiny exercise, and reflect it in their grade boundary recommendations (although, see Section 4.1).

Just for the record, the way in which we use prediction matrices during normal times, i.e. during periods of stability between qualification reforms, should not be described as applying the Comparable Outcomes principle. Although prediction matrices certainly are used to apply the Comparable Outcomes principle – as will be described shortly – they are used differently, here, to achieve a different goal. The use of prediction matrices during normal times is simply a sophisticated approach to operationalising the Similar Cohort Adage, which states that: if the cohort hasn't changed much, then don't expect the pass rate to change much either.

3 Qualification Reforms

If qualifications are to remain relevant to a changing world, they will need to be reformed every so often, perhaps every 5 to 10 years (bearing in mind that some subjects will need reforming more frequently than others). Almost always, qualification reform will involve significant subject content changes. For example, it might be decided that the mathematics curriculum for secondary school students ought to be revised, to include a new component devoted to statistics, and to lose an old component devoted to calculus. Sometimes, qualification reform also involves significant changes to assessment structure or formats. For instance, it might be decided that a qualification that had previously been assessed entirely by coursework ought now to be assessed predominantly by exam.

3.1 Sawtooth Effect

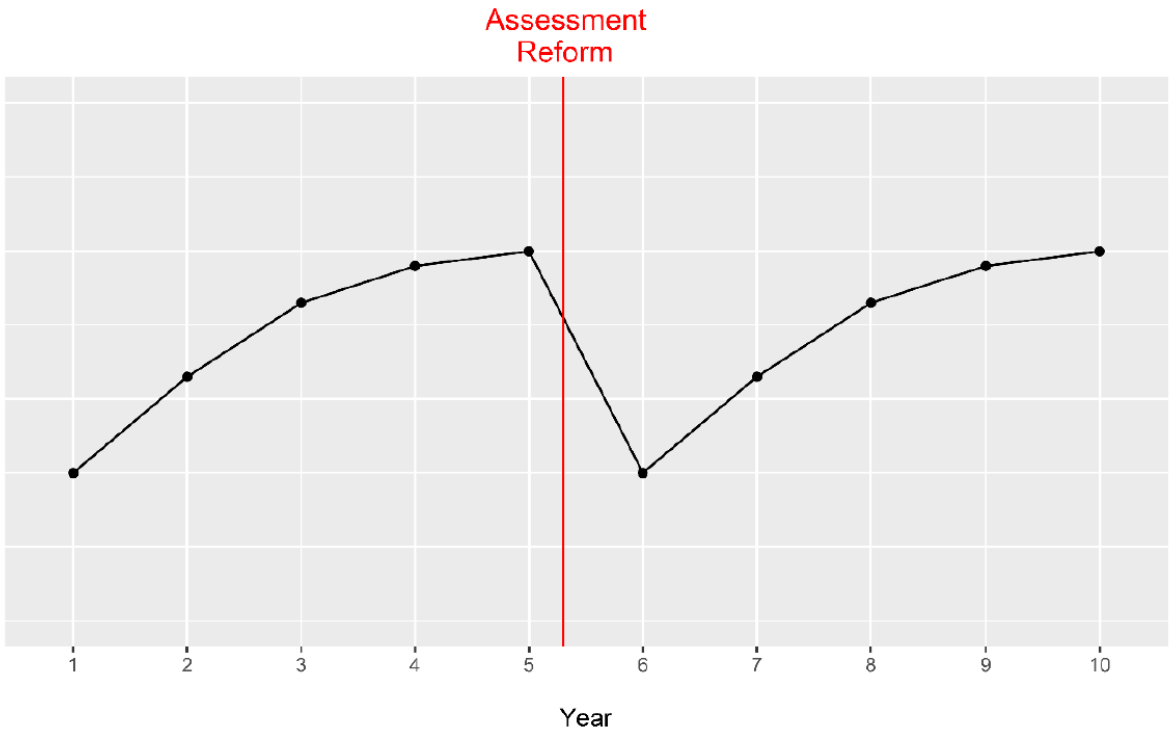

Figure 2. The Sawtooth Effect pattern (adapted slightly from Cuff, 2016, Figure 1).

In the first year following a qualification reform, teachers will face additional challenges: both teaching the new content elements; as well as preparing learners for the new assessment structure/formats. For instance, teachers who had never had to teach statistics would not be able to fall back on a previously developed 'catalogue' of lesson plans. Indeed, they might not be entirely up to scratch on the statistical content themselves. Similarly, teachers and candidates might lack effectively tailored teaching resources, such as textbooks, in the first year of the new syllabus; and there would be no 'past papers' to illustrate the approach to examining the new content elements.

In short, there are good reasons to assume that the first cohort for a reformed qualification will perform less well in their exams than the last cohort pre-reform. As teachers gradually get back up to scratch in terms of teaching the new content elements, and preparing students for the new assessment structure/formats, performances will inevitably begin to rise again. Were we to trace out this likely performance trend, it would look like the tooth of a saw, as illustrated in Figure 2. Hence its name, the **Sawtooth Effect**.

3.2 Maintaining standards across reforms

The Sawtooth Effect presents us with a thorny question. If the performance of the first cohort post-reform is lower than the performance of the last cohort pre-reform, then should the new cohort of students end up with lower grades (at least on average)? After all, it is not really their fault that their performance was lower.

According to the principle of Attainment-Referencing, if the overall level of attainment of the new cohort is lower (on average) than the overall level of attainment of the old cohort, then the new cohort *should* end up with lower grades; even if those candidates were not actually to blame for their lower level of attainment.

Now, it could be said that part of the reason for the post-reform cohort's lower performance was not due to lower attainment, per se; it was simply a matter of lower performance on the exam. In other words, candidates who were assessed via the new exam structure/formats would have been under-prepared for these new demands, and would have been correspondingly less able to demonstrate their true levels of attainment. In other words, being unfamiliar with the new assessment structure and formats, they performed under par. This impairment might persist for a couple of years post-reform; maybe a few. Bear in mind that even during normal times we actively compensate for candidates performing under par on an assessment – for reasons that are solely due to the difficulty of that assessment – by reducing grade boundaries. So, there is a precedent for compensation here, at least.

Yet, there is also a sense in which the performance of the first cohort post-reform genuinely does reflect a lower level of attainment. This might be true as a consequence of teachers in the first year not yet being up to scratch in teaching the new content elements. In other words, in year 1, they would not be as good at teaching statistics as they had become at teaching calculus, having taught calculus for years before the qualification was reformed. If so, then we would expect overall attainment in mathematics, immediately following the reform, to be somewhat lower than immediately prior to it.[6] And, if so, then the principle of Attainment-Referencing would rule that the first cohort ought, indeed, to end up with lower grades.

3.3 Comparable Outcomes

A couple of decades ago, the exams industry in England formalised an approach to dealing with standards across periods of reform.[7] It was decided that, in the exceptional circumstances of the first year of a new qualification:

> candidates taking the new exams should receive, as a group, comparable grades to those which they would have received had they followed the old courses.
>
> (Cresswell, 2003, p.14)

We can think of this as applying a quite different principle for maintaining standards: the **Comparable Outcomes** principle. Of course, this principle sets out a

[6] This would reflect poorer teaching of the statistics component, even if the remaining content were taught just as well as in previous years.

[7] There is evidence of it having been applied previously, albeit informally and inconsistently.

counterfactual: we cannot possibly know how the new cohort would have performed on the old exams, having followed the old syllabus. All that we can do is to estimate this, and we do that using prediction matrices, for each subject exam.

To apply the Comparable Outcomes principle, we ensure that the relationship between prior attainment results and exam grades is the same for the first cohort post-reform as it was for the last cohort pre-reform. This is not the same as awarding an identical distribution of grades. If, for instance, the two cohorts were demographically quite different – let's say that the reformed syllabus attracted many more high-achieving learners – then it would not be right to award exactly the same distribution of grades, pre-reform versus post-reform. We use prediction matrices to adjust for differences in cohort demography, estimated on the basis of prior attainment results. As such, if the two cohorts were:

1. demographically equivalent, then we would engineer their overall grade distributions to be equivalent; but if they were
2. demographically different, then we would engineer their overall grade distributions to reflect the impact of those demographic differences.

This is a different use of prediction matrices from that which occurs during normal times. During normal times, prediction matrices are essentially used to determine a best guess for where each grade boundary ought to lie; with the proviso that this best guess might be revised on the basis of examiner judgement of script evidence. This is to apply the principle of Attainment-Referencing.

> It would be unfair to apply exactly the same attainment standards in the first year following a reform as in the last year preceding it.

To maintain exam standards across periods of reform, the principle of Attainment-Referencing is disapplied, and the Comparable Outcomes principle is applied instead. This states that it would be unfair to apply exactly the same attainment standards in the first year following a reform as in the last year preceding it. Instead, exam standards are carried forward across the transition period purely on the basis of statistical expectations.

We apply the Comparable Outcomes principle to be **fair** to cohorts of candidates who, through no fault of their own, end up achieving lower levels of attainment than previous ones. Expressed like this, it is just as relevant to a pandemic period as to a period of qualification reform. There is a strong argument in favour of applying the Comparable Outcomes principle during Summer 2021, to compensate for learning that has been lost to COVID-19. Unfortunately, whereas all learners will be affected

in a similar way by a qualification reform, the same will not be true for COVID-19. It will not be possible to compensate for learning loss in 2021 where this affects some learners far more than others. There is still a case for applying the Comparable Outcomes principle, albeit only as a partial compensation strategy.

4 Outstanding challenges

There are no perfect solutions to problems related to maintaining exam standards over time; neither between periods of qualification reform – which we like to think of as periods of relative stability – nor across them. Over the decades, we have become much better at addressing such problems. Yet, there are many outstanding challenges to address – both practical and theoretical – and it is possible that some of these challenges may never be solved.

4.1 Judgement versus statistics

There is a widespread misperception that pass rates are always fixed on the basis of statistical expectations nowadays; as though Ofqual required the Comparable Outcomes principle to be applied each and every year, both across periods of qualification reform and between them. This is not true. Nowadays, we require exam boards to apply the Comparable Outcomes principle to bridge a qualification reform, as well as for a short period following it (see below). However, after this short period, we require exam boards to revert to Attainment-Referencing.

The most obvious consequence of applying the Comparable Outcomes principle is that it puts a cap on national pass rates. At least in theory, reverting to Attainment-Referencing removes this cap. Grade boundaries are decided on the basis of a judicious balance between examiner judgement and statistical expectations. This allows for increases or decreases in national pass rates at the subject level – beyond those attributable to demographic change across adjacent cohorts – on the basis of recommendations from grade awarding committees.

In practice, however, if an awarding committee were to recommend grade boundaries that departed substantially from the SRBs for a subject, it would be under considerable pressure to provide a persuasive justification; bearing in mind the fallibility of human judgement. Justifications of this sort can be hard to mount, especially when an awarding committee can find no other basis for supporting their recommendations, beyond their professional judgements of script quality.

A major part of the problem, here, is that confidence in the ability of examiners to judge attainment standards has fallen in recent years; in particular, owing to evidence of continually rising pass rates, but also on the basis of experimental research. Although it is impossible to be definitive, there is a general feeling within the exams industry that certain features of examiner judgement – including a tendency to give students the benefit of the doubt when deciding between equally

plausible adjacent boundaries – may have led to an element of **grade inflation** over the past few decades.[8]

Because of this loss of confidence, there is a tendency nowadays to defer to statistical expectations of grade boundary locations, unless an awarding committee can provide a particularly persuasive justification for not doing so. Unlike in recent decades, the balance has now swung towards greater reliance upon statistical expectations, with less confidence being placed upon examiner judgement. This makes it less likely that pass rates will change over time – other than for reasons to do with the changing demography of a cohort – although change is by no means ruled out.

4.2 Beyond the Disruption Effect

The Sawtooth Effect describes a consequence of qualification reform, whereby candidates' performances suddenly dip across the transition from old to new, followed by a gradual rise back up again over a number of years. Clearly, then, the Sawtooth Effect can be deconstructed into two separate effects: a **Disruption Effect**, where performance suddenly falls; and an **Enhancement Effect**, where performance gradually rises back up again. We apply the Comparable Outcomes principle to compensate for the Disruption Effect, but what about the Enhancement Effect?

After all, if we that think it is unfair for candidates to be penalised for their teachers not being up to scratch in the first year post-reform, then is it not equally unfair for candidates to benefit from their teachers gradually returning back to form in the second and third years post-reform?

In other words, if it is fair to apply the Comparable Outcomes principle in the first year post-reform (so the first cohort is not unduly penalised), then is it not equally fair to apply the Comparable Outcomes principle in the second, or third year post-reform (so that the second, or third cohorts are not unduly rewarded)? The argument is compelling.

The problem, here, is exactly how long we ought to continue applying the Comparable Outcomes principle, before returning to Attainment-Referencing. We do have some evidence to suggest that the Sawtooth Effect may last for the 'first few years' following a qualification reform (Cuff, 2016). However, this comes from just a single exploratory study, and its conclusions are far from definitive. The current policy is for Comparable Outcomes to be applied for a couple of years, after which it is assumed that Attainment-Referencing ought to come into play again.

[8] This is an understandable reaction, of course, when forced to chose between two equally plausible marks. The problem is its potential for creating a ratcheting effect, over a period of decades.

4.3 Beyond the Recovery Effect

The label Sawtooth Effect was coined in North America, where it has been associated most strongly with concerns over score inflation in accountability testing, which parallels the idea of grade inflation in certification examining. It has been most extensively discussed, and theorised, by Koretz (2008; 2017), who has identified 7 different types of test preparation, i.e. 7 different reasons why results on accountability tests might rise over time. The first 3 of these reasons – teachers working more effectively, teachers teaching more, and teachers working harder – are entirely legitimate. They are exactly what policy makers would hope to achieve under incentives for school improvement. Conversely, his seventh reason – teachers cheating more – would be entirely illegitimate. This is not a matter of successive cohorts learning more or learning better. Gains attributable to cheating have no basis in rising attainment, they are purely performance gains; and highly unethical performance gains, at that.

Of most interest to the present discussion are his remaining reasons – alignment, reallocation, and coaching – which tend to figure prominently in discussions of the Sawtooth Effect in the USA. Each of these 3 is a likely cause of rising results on accountability tests over time. Yet each, in its own way, is **contestable**.

Table 1 presents a subtle reconfiguration of these concepts, adapted to the context of examining in England. It incorporates our own thinking on the Sawtooth Effect, discussed earlier, which emphasises the idea of teachers getting back up to scratch post-reform. It identifies 4 potential causes of the Enhancement Effect, each of which provides a reason why the performance of successive cohorts of candidates might gradually rise over time. Importantly, each of these causes is of contestable significance for one reason or another.

Effect	Mechanism	Impact
Realignment	Teachers become better at teaching new content elements	Attainment gain (authentic/unimportant)
Adeptness	Teachers become better at preparing learners to respond to new assessment structure/formats	Performance gain (inauthentic)
Coaching	Teachers begin to identify and pass on to their students hacks & strategies for scoring more marks than they deserve	Performance gain (inauthentic)
Reallocation	Teachers become better at question spotting, reallocating their instructional resources towards the frequently tested content elements/assessment formats	Performance gain (inauthentic)

Table 1. Potential causes of the Enhancement Effect

The first two causes have already been discussed. They concern teachers becoming better at teaching new content elements (Realignment), and teachers becoming better at at preparing learners for new assessment structures and formats (Adeptness).

We use the concept of **Realignment** to characterise an unimportant – albeit in a limited sense authentic – performance gain. These are gains in cohort performance, over the first few years of a reformed syllabus, which are attributable simply to teachers becoming better at teaching the new content elements. They are authentic, in a sense, because these gains really do reflect teachers becoming better at teaching those new elements, and therefore learners learning them better. However, they are unimportant because they are specific to those new content elements, and they arise from an artificially low baseline. They reflect better teaching of the new content elements, but not better teaching, per se. In short, they are nothing to boast about (neither for a teacher nor for a politician).

This is a subtle, but critical, point. By analogy, imagine that you sell your car tomorrow and buy a van. It will take you a few weeks to get used to driving it. But would you say that you had suddenly become a significantly worse driver? And, a few weeks later, would you then say that you had now become a significantly better one? If you were being really picky, you might say that, for a few weeks, you were a significantly worse van driver than car driver; but that you soon became just as good a van driver. But why be so picky? This is not an important improvement in general driving ability that would be worthy of boasting about.[9] It is an unimportant improvement related to a specific vehicle.

In a similar way, we use the concept of **Adeptness** to characterise an inauthentic (and therefore also unimportant) performance gain. These are gains in cohort performance, over the first few years of reformed assessment arrangements, which are attributable solely to teachers becoming better at preparing learners for the new assessment structure/formats. The critical distinction, here, is between the content that is being assessed and the process by which it is assessed.

For the sake of clarity, let's assume that the reformed qualification assesses the same content via a new assessment format; for example, a reading comprehension exam changes from short-answer format to multiple choice format. Unused to this new multiple choice format, in the first year of the new exam, teachers neglect to train students always to provide an answer, even if it is simply a guess. After a year

[9] Compare this, for instance, with having practised for, and then passed, an advanced driver course.

or so, most teachers have realised that a substantial number of students have lost out on marks that they might have achieved by guesswork alone. They revise their exam preparation to emphasise the importance of guessing rather than leaving answers blank. Ultimately, this translates into a small boost in the performance of the national cohort over the first few years. But this boost reflects *nothing more* than better performance on the exam; it does not correspond to an underlying change in cohort attainment.

Coaching and reallocation come from the North American literature, but translate directly into the context of examining in England. They both rely upon the idea of growing familiarity with the exam in question: with experience from repeated administrations, ways to game the exam reveal themselves.

Coaching refers to training in techniques that allow candidates to score high marks on assessment tasks without a high level of ability in the content element being examined. This might involve the teacher spotting a blatant hack; for example, they might notice that correct options within multiple choice questions tend to include more words than incorrect options, or are almost never the third option. Alternatively, coaching might involve a more subtle strategy; for example, training candidates to routinely reproduce features that are consistently rewarded by mark schemes, without actually exercising the skills that those features are supposedly indicative of.

Reallocation is more subtle, again; but no less problematic. It refers to teachers becoming increasingly aware of content elements that are predictably absent from the exam. As it becomes increasingly evident that they are never examined, it becomes increasingly evident that they do not need to be taught. Teaching time is reallocated to content elements that appear more frequently on the exam. Over time, attainment on the predictably examined content rises, but at the expense of falling attainment on the predictably not examined content. This translates into no overall rise in attainment, which is why the observed performance gain (across the examined subset of the domain) has to be considered inauthentic.

These four causes can be split in two, according to how soon their effects are likely to be felt. Both Realignment and Adeptness are likely to start straight away, following a qualification reform. In addition, because teachers will want to get back up to scratch as soon as possible, it seems likely that they will also peter out fairly soon; perhaps after just a few years. Together, these causes may explain the first part of the Enhancement Effect, which we might describe as the **Recovery Effect**.

Coaching and Reallocation, on the other hand, seem likely to start only after a number of years has elapsed. This is because they are premised upon the idea of teachers becoming increasingly familiar with the foibles of the new exam; and this can only happen after a number of administrations. Similarly, as word of such foibles spreads, we might assume that the effects of Coaching and Reallocation may become more pronounced over time. Together, they may explain the second part of

the Enhancement Effect, which we might describe as the **Augmentation Effect**. This deconstruction of the Sawtooth Effect is illustrated in Figure 3.

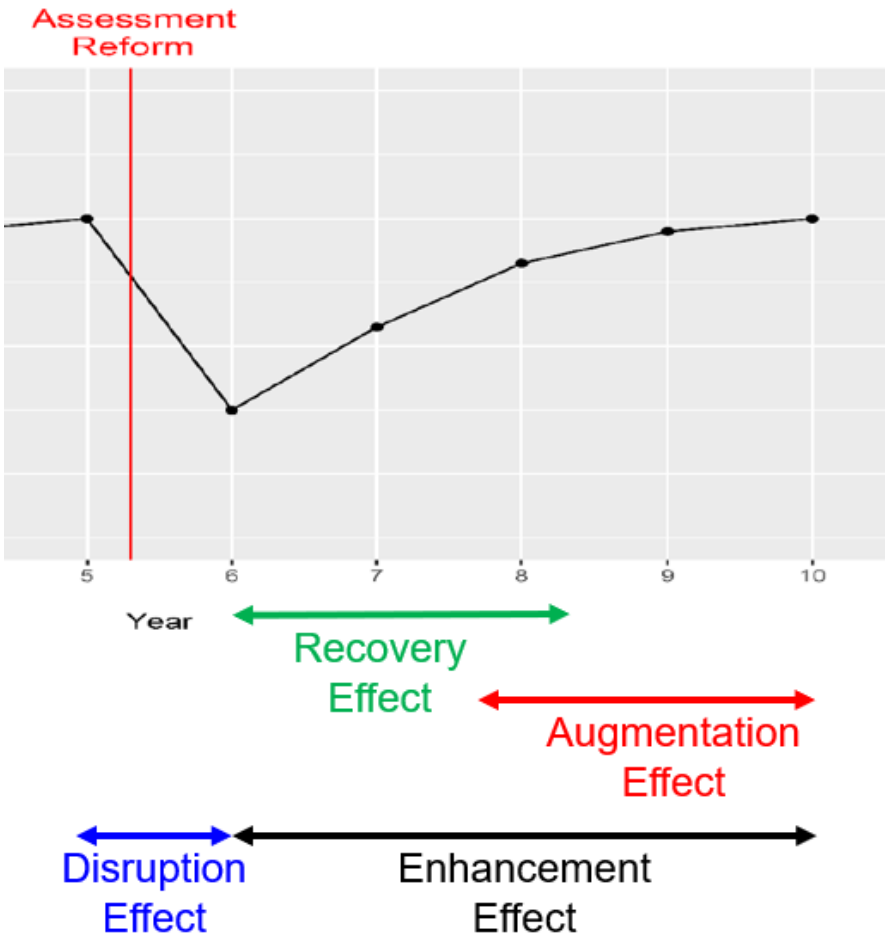

Figure 3. A deconstruction of the Sawtooth Effect

This deconstruction raises an important, but as yet unanswered, question concerning optimal strategies for tackling the Sawtooth Effect. Nowadays, we always apply the Comparable Outcomes principle in the first year following a qualification reform, to counteract the Disruption Effect (so as not to penalise the first cohort). As we have seen, there is also a strong argument for continuing to apply the Comparable Outcomes principle for a number of years following the qualification reform, to counteract the Recovery Effect (so as not to unduly reward subsequent cohorts).

But what, if anything, can be done about the Augmentation Effect? It is not at all clear. If, for instance, we were to continue to apply the Comparable Outcomes principle indefinitely, then this would certainly achieve one important end – it would rule out the possibility of grade inflation. However, it would also rule out the possibility of ever recognising any authentic rises in attainment over time.

One argument against attempting to compensate for the Augmentation Effect by applying the Comparable Outcomes principle is that it is unlikely to operate consistently across schools. It seems now to be generally recognised that the pressure of accountability will lead some teachers in some schools to make full use of gaming techniques. Importantly, though, many teachers in many schools will not.

Comparable Outcomes is a great tool for dealing with factors that can be assumed to affect all candidates in a similar way; for example, when an exam paper proves to be particularly easy or hard; or when teachers need to get back up to scratch with new content elements and new assessment structure/formats as soon as possible. However, when factors operate differently across candidates in the cohort, the Comparable Outcomes principle cannot adequately compensate for them, because it can only apply a uniform adjustment. More importantly, if grade boundaries were raised to accommodate the impact of a minority of schools using gaming techniques, this would subtly penalise candidates from the majority of schools that refrained from such practices.[10]

5 Conclusion

Producing successive versions of an exam according to a common design blueprint is a necessary step, but not a sufficient one, to ensure that each new version applies exactly the same standard as the one that preceded it. It is impossible to guarantee, from inspection alone, that each new version has exactly the same level of difficulty. Standards for each new exam therefore need to be fine tuned – by deciding whether to locate its grade boundaries a little higher, a little lower, or in the same place – in the light of candidates' actual performances.

There is a science to this process of fine-tuning exam standards – which is known as the process of grade awarding – but this is not an absolutely precise science. It requires a judicious balance of examiner judgement and statistical expectations.

Even during normal times – when curriculum, syllabus, and assessment arrangements do not change – the grade awarding process is potentially error prone.

---

[10] This is effectively a double-whammy. Imagine a candidate from a non-gaming school whose level of attainment (and quality of performance) landed them right where the A* grade boundary ought to be. Now, imagine a second candidate whose level of attainment should have landed them just below where the A* grade boundary ought to be. As it happens, this is year 3 of the reformed qualification, and teachers (including the second candidate's teacher) have suddenly realised how to game the exam. This raised the second candidate's quality of performance above that of the first. Now, because many candidates had their performances inflated artificially, applying the Comparable Outcomes principle meant increasing the A* grade boundary by 1 mark. This meant that the first candidate ended up being awarded an A, despite their level of attainment genuinely warranting an A*. In other words, they lost out in comparison to candidates within the same cohort (whose teachers gamed the system) and they lost out in comparison to candidates from the previous cohort (whose teachers did not game the system).

In recent years, we have become increasingly wary of placing too much confidence in examiner judgement; which means that we have come to place more weight on statistical expectations. Statistical expectations, unfortunately, do not help us to track potentially small, albeit authentic and important gains (or drops) in cohort attainment over time. Ofqual is currently researching the potential for increasing the rigour of judgemental techniques for maintaining exam standards, and for placing more confidence in their outcomes.

When qualifications are reformed – and curriculum, syllabus, and assessment arrangements do change – a different kind of challenge arises. During the transition year itself, we would generally predict a small but significant drop in cohort attainment. However, it is not the fault of the new cohort that the qualification has been reformed. So, we are reticent to allow their distribution of grades to fall. Instead, we apply the Comparable Outcomes principle – a fairness principle – to ensure that the new cohort is not penalised. We engineer pass rates to ensure – to the best of our ability – that they achieve the same distribution of grades that they would have achieved, had they studied under the old arrangements, and taken the old exam.

By the second year after a qualification reform, and into the third, teachers will generally have got the hang of teaching the new content elements, and of preparing learners for the new assessment structure/formats. We have called this the Recovery Effect. However, just as the worse performance of the first cohort post-reform was not their fault, nor is the better performance of the second and third cohort necessarily to their credit. At least some, and potentially all, of this performance gain can be described as either inauthentic or unimportant. Under these circumstances, there are good reasons to continue to apply the Comparable Outcomes principle for a couple of years post-reform.

Ideally, we should disapply the Comparable Outcomes principle as soon as possible, and return to Attainment-Referencing. In theory, at least, this should mean that we could begin to track potentially small, albeit authentic and important gains (or drops) in cohort attainment over time. We do, in fact, shift back to Attainment-Referencing after a couple of years post-reform.

Unfortunately, in practice, there is still a risk that pass rates may become inflated over time; that is, a risk that they are corrupted by inauthentic and/or unimportant performance gains. A key concern, here, is the Augmentation Effect, which appears not to be something that we could easily factor out of exam results.

Trends in exam pass rates over time are notoriously difficult to interpret under the best of circumstances. Even when it looks as though exam results are improving over time, this is often a simple consequence of the demography of the cohort having changed. In other words, the subject is neither being taught better nor learned better than before, it is simply being studied by a stronger cohort. Threats such as Coaching and Realignment, which give rise to the Augmentation Effect,

make these trend lines even harder to interpret definitively. In short, we must always exercise an element of caution when interpreting trends in exam results over time.

6 References

Centre for Education Research and Policy (undated). *A basic guide to standard setting. Version 1.5.* Manchester: Assessment and Qualifications Alliance.

Coe, R. (2013). *Improving Education: A triumph of hope over experience*. Inaugural Lecture of Professor Robert Coe, Durham University. 18 June.

Cresswell, M.J. (2003). *Heaps, Prototypes and Ethics: The consequences of using judgements of student performance to set examination standards in a time of change*. London: University of London Institute of Education.

Crofts, J.M. & Caradog Jones, D. (1928). *Secondary School Examination Statistics*. London: Longmans.

Cuff, B.M.P. (2016). *An Investigation into the 'Sawtooth Effect' in GCSE and AS / A level Assessments*. Coventry: Office of Qualifications and Exams Regulation.

Koretz, D. (2008). *Measuring Up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Koretz, D. (2017). *The Testing Charade: Pretending to make schools better*. Chicago, IL: The University of Chicago Press.

Newton, P.E. (2020). *What is the Sawtooth Effect? The nature and management of impacts from syllabus, assessment, and curriculum transitions in England*. Coventry: Office of Qualifications and Exams Regulation.

Taylor, R. & Opposs, D. (2018). Standard setting in England: A levels. In J. Baird, T. Isaacs, D. Opposs and L. Gray (Eds.). *Examination Standards: How measures and meanings differ around the world* (pp.100-113). London: Trentham Books.

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344

public.enquiries@ofqual.gov.uk

www.gov.uk/ofqual

November 2020

Ofqual/20/6710/2