RESEARCH AND ANALYSIS

What is the Sawtooth Effect?

The nature and management of impacts from syllabus, assessment, and curriculum transitions in England

ofqual

Authors

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

Contents

1 Preface

This report is about the interpretation of assessment results, and the steps that assessment agencies can take – either before or after awarding those results – to ensure that they are interpreted appropriately; or, at least, to minimise the likelihood of misinterpretation. Some of the issues at stake are quite subtle, and complicated to unpick. This report attempts to explain why, introducing a variety of concepts that can be used to tease them apart, and to explore their implications. It begins by introducing some of the core ideas underpinning the subsequent analysis.

Philosophy

> In judging that's someone's performance is or is not intelligent, we have, as has been said, in a certain manner to look beyond the performance itself. […]
>
> We observe, for example, a soldier scoring a bull's eye. Was it luck or was it skill? If he has the skill, then he can get on or near the bull's eye again, even if the wind strengthens, the range alters and the target moves. Or if his second shot is an outer, his third, fourth and fifth shots will probably creep nearer and nearer to the bull's eye. He generally checks his breathing before pulling the trigger, as he did on this occasion; he is ready to advise his neighbour what allowances to make for refraction, wind, etc. […]
>
> To decide whether his bull's eye was a fluke or a good shot, we need […] to take into account more than this one success. Namely, we should take into account his subsequent shots, his past record, his explanations or excuses, the advice he gave to his neighbour and a host of other claims of various sorts. There is no one signal of a man's knowing how to shoot, but a modest assemblage of heterogenous performances generally suffices to establish beyond reasonable doubt whether he knows how to shoot or not.

(Ryle, 1949, p.45)

In this passage, the philosopher, Gilbert Ryle, captures both the nature and the challenge of educational assessment. A constant feature of everyday life is that we attribute degrees of knowledge, skill, and understanding to those with whom we interact; in order to help us to manage those interactions. Educational assessment is essentially a formalisation of this everyday attributional process, in relation to broad domains of knowledge, skill and understanding. Through educational assessment, we aim to measure the overall level of knowledge, skill and understanding attained by a learner, by the end of their course of instruction in a domain of learning; and we look to exam results, e.g. a grade A or a grade C in A level geography, to tell us the degree to which learners have mastered the subject domain in question.

In terms of the challenge of educational assessment, Ryle observes that even ostensibly straightforward proficiencies, like the ability to shoot a bull's eye, involve a complex set of skills, manifested in a variety of ways. He explains how we attribute competence on the basis of performance; but ideally not on the basis of particular performances. This is because any particular performance might be a fluke, or a flunk; which means that we can only (properly) determine a person's level of competence by taking into account a range of performances. The challenge of attributing competence on the basis of performance is at the heart of educational assessment, and it is fundamental to understanding the Sawtooth Effect.

Attainment

Educational assessment generally involves measuring a broad domain of knowledge, skill and understanding; typically, by sampling from the elements that comprise it. We tend to describe the resulting measurement of knowledge, skill and understanding using one of a variety of essentially equivalent terms; e.g. a level of competence, a level of proficiency, a level of achievement, or a level of attainment. For the purpose of the present report, we will refer to 'attainment' throughout. Thus, the distinction at the heart of this report is between **performance** and **attainment**.

As suggested in the preceding section, the distinction between performance and attainment is between: particular versus general; and observed versus attributed. In other words, a performance is a particular instance of behaviour, which is observed on a particular occasion; whereas attainment is a general proficiency, which is (properly) attributed on the basis of numerous behaviours in a variety of contexts.[1] For instance, when we say that an individual has attained a certain learning outcome, e.g. mastered addition, we are attributing to them a general proficiency in adding numbers, rather than observing that they have solved a particular set of addition problems. The idea of a certain *level* of attainment generalises this idea one step further, referring to an individual's overall mastery of a subject area, such as geography, or physics.

According to this analysis, a level of attainment is (properly) attributed on the basis of numerous performances across a variety of contexts. The greater the number and variety of performances we are able to observe from an individual, the more confident we can be in attributing a certain level of attainment to them. Having observed numerous performances from an individual, across a variety of contexts, we would tend *not* to attribute to them the level of attainment that corresponded to their highest quality performance (possible fluke), *nor* to their lowest quality performance (possible flunk). Instead, we would attribute to them a level of

[1] A broad, encompassing sense of 'behaviour' is intended here; to include actions, utterances, pieces of writing, and so on.

attainment commensurate with how they tended to perform, in general, across the board; explicitly recognising that the quality of their performance on any particular occasion might well be misleading, for all sorts of different reasons.

Although results from educational assessments, such as GCSE grades, are based upon performances – typically, one-off exam performances – they are intended to be interpreted more generally in terms of attainment, and to be bestowed with more significance than simply how a candidate happened to have performed on a particular day in a particular exam. Interpreting an exam result is therefore a matter of **generalising** from the quality of performance that we observe in the exam context to the level of attainment that would properly be attributed to the individual in question, were they to be observed across a large number and wide variety of performances in non-exam contexts.[2] That is, we interpret the exam result in terms of the candidate's *general* proficiency – their overall level of attainment in GCSE physics, let's say – and *not simply* in terms of how they happened to have performed in their physics exam, one day in June.

We interpret results like this because of the **uses** to which they are put; that is, we *use* results *as though they do* say something more general about students' levels of attainment. This is exactly the presumption when, for instance, a GCSE physics grade 5 is specified as a minimum requirement for being admitted onto an A level physics course. We want to be reassured that the learner has a satisfactory foundation for commencing the higher course of study. The last thing that we would actually *want* to do, in this situation, would be to reward them with a place on the A level course, as some kind of perverse prize for having fluked a grade 5.[3]

Fallibility

Educational assessment is inherently fallible. We cannot always be confident in inferring attainment on the basis of successful performance; and we cannot always be confident in inferring lack of attainment on the basis of unsuccessful performance.

For instance, answers to questions in certain formats – including multiple-choice and yes/no formats – can straightforwardly be guessed. A correct guess is tantamount to a fluked bull's eye. Indeed, the chances of being able to pass a 10-item yes/no-format test by guessing alone are far from trivial. But there are other reasons why a candidate's test performance might diverge wildly from their level of attainment; for

[2] Importantly, exams are specifically designed to render this generalisation as legitimate as possible. The inferences involved in generalising from an exam performance to a real-world proficiency are at the heart of Michael Kane's argument-based approach to validation (e.g. Kane, 2006; 2013).

[3] Of course, in the absence of any reliable, independent evidence concerning their level of attainment, we might still *end up* doing this! But, if they really had fluked a grade 5, and their actual level of attainment was far more like a grade 3, then we would not actually *want* to admit them.

example, they might simply cheat. Equally, though, performance and attainment might diverge in the opposite direction for candidates who fail to put any effort into completing the test, or whose performance is greatly inhibited by test anxiety.

Guessing, motivation, and anxiety are all candidate-related explanations for a lack of convergence between performance and attainment. But there are test-related explanations too. Ryle observed that a competent soldier is likely to get close to the bull's eye even if the wind strengthens, the range alters, and the target moves. This recognises the fact that some contexts of performance will be more demanding than others; which seems likely to be true of all competences. In a testing context, if the test only happens to sample from the more demanding contexts, then it is likely to result in an under-estimate of the candidate's level of attainment. For instance, we might under-estimate a learner's driving proficiency (their level of attainment in driving) were we simply to observe them driving in storm conditions. Or, we might over-estimate their driving proficiency, were we simply to observe them driving in a town with no hills, roundabouts, or traffic lights, on a quiet day.

Understanding the many different reasons why (quality of) performance cannot necessarily be interpreted *directly* in terms of (level of) attainment provides a useful foundation understanding the Sawtooth Effect.

Standards

The distinction between performance and attainment is also critical to understanding the maintenance of assessment standards. GCSE and A level awarding bodies manage the maintenance of exam standards when they locate grade boundaries. A grade boundary is the lowest total mark (achieved on the exam overall) that is judged to be worthy of a particular grade. So, the GCSE grade 7 boundary is the lowest total mark judged to be worthy of a grade 7, for the exam in question. Even when successive exams – e.g. GCSE physics 2018 versus GCSE physics 2019 – are built according to the same blueprint, with the same maximum mark, their grade boundaries may be located at different marks. In a nutshell, this is because a mark corresponds to a certain quality of performance, whereas a grade boundary is intended to correspond to a certain level of attainment, and the two do not always align.

The idea, here, is that the GCSE grade 7 boundary mark should always correspond to exactly the same level of attainment, from one exam to the next. But, just like in the driving proficiency example above, the context within which candidates demonstrate their levels of attainment – that is, the particular questions that appear in each exam – may affect the quality of the performances that we observe. If, for instance, the particular set of physics questions selected for 2019 happened to be especially difficult, then we might expect the quality of candidates' performances (and therefore their marks) to drop, *even if the average level of attainment of the*

2019 cohort was exactly the same as the average level of attainment of the 2018 cohort. This is why awarding bodies manage the maintenance of exam standards by deciding whether to locate each grade boundary either higher, lower, or at the same mark as it was located in the previous year. If the particular set of physics questions selected for 2019 happened to be more difficult than the set of questions selected for 2018, then the 2019 GCSE physics grade 7 boundary mark would be located correspondingly lower. If the 2019 exam were judged to be 2-marks-worth harder, then the grade 7 boundary would be located 2 marks lower.

This is the **traditional logic** of **grade awarding**. Grade awarding therefore refers to a step that awarding bodies take, *before awarding results*, to ensure that those results can be interpreted appropriately; that is, to ensure that they can be interpreted in terms of an equivalent level of attainment, from one year to the next.

Transitions

We have just considered the traditional logic of grade awarding, which is generally presumed to apply during periods that are characterised by stability. The present paper, and the Sawtooth Effect more generally, is concerned with what happens during periods of transition; for instance, in the wake of a reform of the national curriculum and associated assessment arrangements.

The General Certificate of Secondary Education (GCSE) was introduced for first teaching in 1986 and was first examined in 1988. It represented a huge transition in syllabus, assessment, and curriculum arrangements. This included the abolition of a dual-route certification model (O level alongside CSE), the introduction of new syllabuses based upon National Criteria, and the widespread incorporation of coursework into the assessment process (which had previously been associated mainly with CSE).

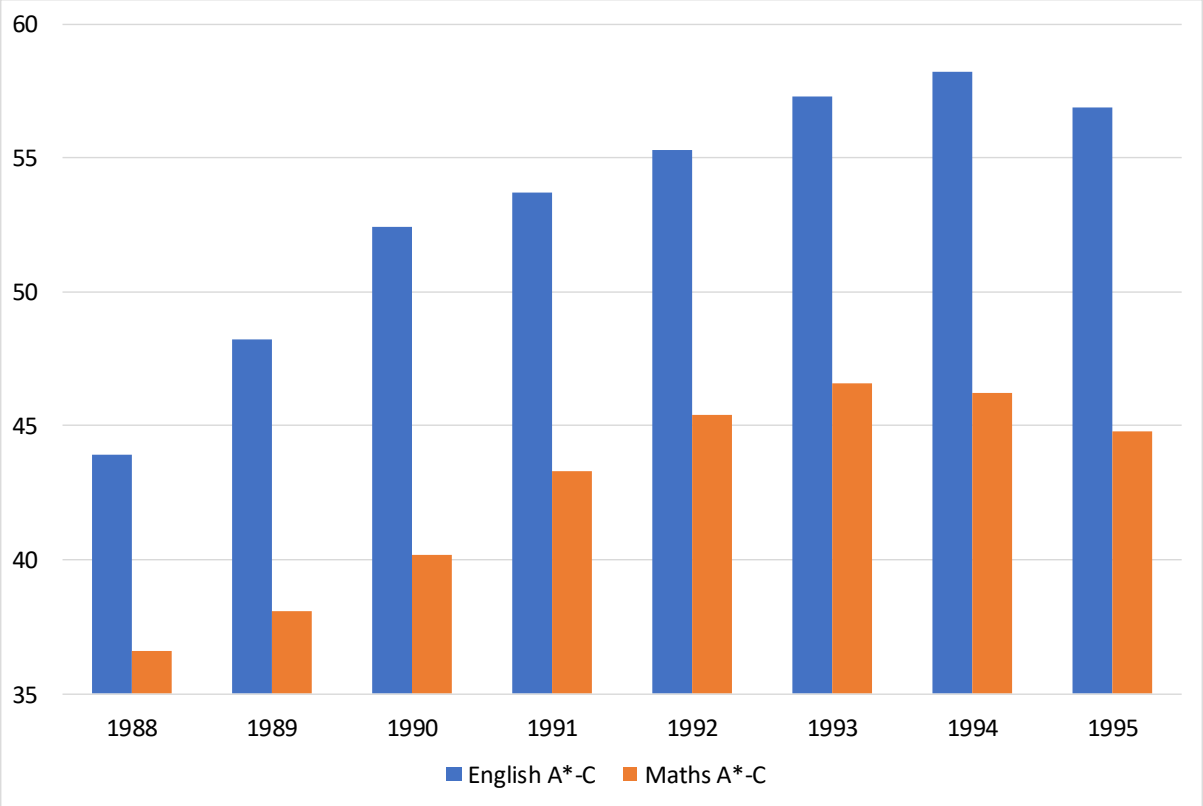

Figure 1. Cumulative percentage of GCSE subject results at grade C and above[4]

In stark contrast to O level exams, where the cumulative percentages of candidates at each grade had tended to remain fairly similar over time, Figure 1 illustrates how the cumulative percentages of candidates at GCSE grades rose steadily during the early years of the new qualification. We can see that the cumulative percentage of candidates who were awarded grade C or above in English rose from 44% in 1988 to 58% in 1994, where the rise appears to tail off slightly. Similarly, the cumulative percentage of candidates who were awarded grade C or above in maths rose from 37% in 1988 to 47% in 1993, where it appears to tail off. But what ought we to make of this steady rise in results during the early years of the GCSE?

Well, assuming that awarding bodies were intending to apply the traditional logic of grade awarding, as described above – which they would have claimed to be doing at the time – and assuming that they were generally successful in doing so, then we ought to be able to infer that levels of attainment in maths and English were rising in England and Wales; and quite substantially so.

[4] This included grade A* from 1994. These data were originally prepared by the Joint Council for the GCSE, although this subset was collated by Smithers (2017). The data relate to England and Wales only.

In recent years, however, we have become far more sceptical of cohort-level changes in performance, from one year to the next, following transitions in syllabus, assessment, and curriculum arrangements. From this more sceptical perspective, we would ask whether we should necessarily interpret cohort-level rises in *performance* as though they represented cohort-level rises in *attainment* (even assuming that fluctuations in exam difficulty had effectively been managed). And even if we were prepared to interpret them as cohort-level rises in attainment, we might still ask whether there might be grounds for questioning the significance of those rises.

In short, this more sceptical perspective would encourage us to explore the possibility that levels of attainment in maths and English *may not* have risen substantially between 1988 and 1993/4, despite what the trends lines appeared to suggest. This is to introduce the Sawtooth Effect.

2 Background

Over the past few years, the Sawtooth Effect has increasingly featured in discussions concerning the maintenance of standards over time within GCSEs, A levels, and other regulated assessments and qualifications in the UK. But what exactly do we mean by the Sawtooth Effect? How does it occur? Why does it occur? And what should we do about it? Although none of these questions is entirely straightforward, the present report attempts to provide some answers.

At Ofqual, we tend to describe the Sawtooth Effect as though it were an effect upon assessment performances – during the early years of a reformed qualification – attributable to an initial lack of familiarity with the new form of that qualification, e.g.

> We know from our research on the sawtooth effect, that student performance dips a little in the first years of a new qualification, because teachers are less familiar with the new specifications, and there are fewer support materials and past papers for students to use.

The implication, here, is that we should *expect* the quality of candidates' performances to drop – from the last year of a pre-reformed qualification to the first year of its reformed counterpart – owing to their teachers' initial lack of familiarity with the new form. Consequently, we should also *expect* the quality of candidates' performances to rise gradually, over the next few years, as their teachers become increasingly familiar with it. This hypothetical pattern, which resembles the tooth of a saw, is illustrated in Figure 2.

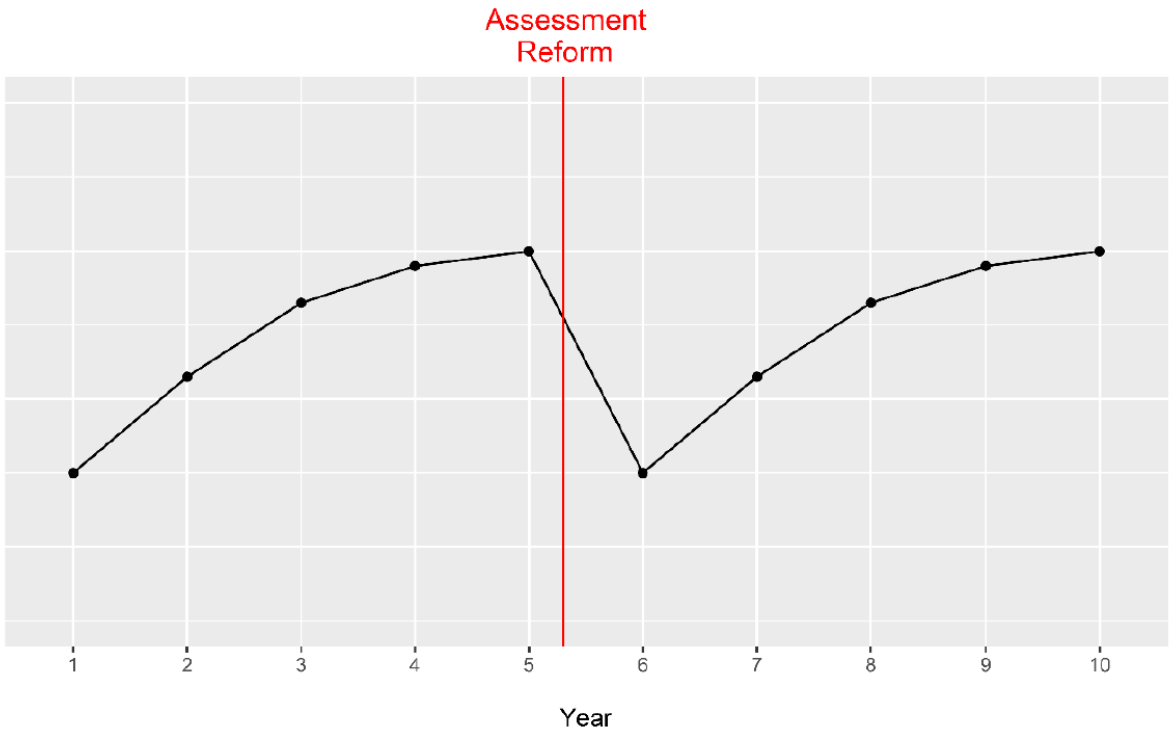

Figure 2. The Sawtooth Effect pattern (adapted slightly from Cuff, 2016, Figure 1).

This pattern of sudden drop followed by gradual rise is reflected in the following explanation of the Sawtooth Effect:

> The 'Sawtooth Effect' is where cohort performance on high-stakes assessments drops after assessment reform, and then improves over time as test familiarity increases.
>
> (Cuff, Meadows, & Black, 2019, p.321)

Figure 2 is adapted from a recent Ofqual publication, which reported upon an impressive body of evidence concerning the occurrence of Sawtooth Effects in GCSE and AS/A level assessments (Cuff, 2016). In the original version of this graph, its Y axis was labelled 'outcomes' and (as in the present version) the transition point was labelled 'assessment reform'. However, both of these labels beg complicated questions, which the present paper will consider in more depth. For instance, is the Sawtooth Effect basically concerned with trends in performance, or with trends in results, or perhaps with other kinds of trends as well (or instead)? Similarly, is the Sawtooth Effect basically concerned with assessment transitions, or with syllabus transitions, or perhaps with other kinds of transitions as well (or instead)? To answer questions like these, we need to consider carefully the possible causes of the effects that we are focusing upon. For instance, are these purely familiarity effects, or are there other kinds of effect at work here too?

As will become clear, the Sawtooth Effect – as the term has come to be used in the UK – is a slightly 'fuzzy' concept, which embraces a number of separable effects, and which hints at a variety of causes. The following analysis introduces the more general concept of **transition impacts** – impacts upon candidates (upon their performances and their attainments) arising from syllabus, assessment, and curriculum transitions – to describe the broader conceptual landscape that tends to be brought into view when discussing the Sawtooth Effect.

3 The nature of transition impacts

The following sections attempt to unpack what we (in the UK) have come to mean by the Sawtooth Effect, as applied to regulated qualifications and assessments. After considering the origins of this term in the USA, and discussion of analogous phenomena in the UK, we will explore issues of particular relevance to the UK context, via a more general consideration of the nature of transition impacts.

USA Origin

The Sawtooth Effect entered the educational assessment literature during the 1990s, following seminal work by Robert Linn and colleagues (e.g. Linn, Graue, & Sanders, 1990; Shepard, 1990; Koretz, et al., 1991; Shepard, 1997; Linn, 2000). They were writing in a North American context, in which nationally standardised achievement tests – e.g. tests of attainment in maths and reading, developed by a plethora of commercial test publishers – were being used widely for accountability purposes. More specifically, the term was introduced in the context of a national debate concerning how it could possibly be true that all 50 states were performing 'above the national average' on such tests (Cannell, 1988).[5]

In response to this strange phenomenon, Linn, et al (1990) reflected on the pattern that could often be observed in test result trend lines when state education departments changed the tests that they were using for accountability purposes. In the classic example, a state would administer exactly the same suite of tests from one year to the next, to students from a particular year group, e.g. grade 5. Over a period of years, the mean percentile rank achieved by successive grade 5 cohorts would increase, *against nationally standardised norms for that suite of tests*. The state would then transition to a new suite of tests; that is, to tests developed by a different publisher, albeit assessing the same subject areas, e.g. maths and reading. And the next grade 5 cohort would achieve a significantly lower mean percentile rank, *against nationally standardised norms for the new suite of tests*. Once again, though, over a period of years following the transition, the mean percentile rank achieved by successive grade 5 cohorts (on the new suite of tests) would increase. The pattern of results thus created, evident from trends in mean percentile ranks over time, resembled the tooth of a saw – hence, the Sawtooth Effect.

Various explanations have been proposed for this effect, but a particularly interesting and important one relates to the idea of teachers and students becoming increasingly well-versed in the assessment process itself. This hypothesis proposes:

[5] The Sawtooth Effect, and a broader body of work from the USA on Score Inflation, is discussed comprehensively in two books by Dan Koretz: *Measuring Up* (Koretz, 2008); and *The Testing Charade* (Koretz, 2017).

that the improvement in test scores over time is primarily due to successive cohorts becoming better at tackling the *kinds of tasks* that appear in the test, owing to increasing test/task familiarity; and, therefore, that this improvement will not generalise to related tasks in a new test (with a slightly different format whilst covering exactly the same content domain) because the improvements do not represent **robust understanding**, i.e. **valuable learning**. In other words, whilst rising results represent an improvement in cohort *performance*, they do not represent an improvement in cohort *attainment*. This hypothesis has been presumed to have particular force when tests are used for accountability purposes, as this creates a situation in which teachers face a strong (perverse) incentive to improve their students' test scores.

Supporting this conclusion, Lorrie Shepard (1997) presented evidence that demonstrated how students who had been prepared for state accountability tests failed to generalise the ability to solve questions posed in a familiar format to questions posed in an unfamiliar format. This contrasted starkly with the performance of an equating sample – who had not been prepared in the context of accountability testing – whose ability did generalise from familiar to unfamiliar question formats. Perhaps, then, in the context of accountability testing, their teachers had become better and better at circumventing assessment demands, by identifying hacks and strategies that lead to correct performances on familiar tasks, but that do not lead to robust understanding/valuable learning. By **coaching**, i.e. by preparing their students to be able to apply these hacks and strategies, they were able to yield gains in test scores over time which did not correspond to authentic learning gains.

This most radical interpretation of the Sawtooth Effect proposes that gains on standardised tests used for accountability purposes are *entirely* spurious: students in successive cohorts demonstrate superior *performances* on the kinds of tasks that appear in the tests – with which their teachers have become increasingly familiar over time – yet this superior performance cannot be equated with superior *attainment*, since it does not generalise to related tasks. In other words, this is *purely* a test/task familiarity effect.

An alternative, or complementary, interpretation is that gains on such tests are at least partly authentic, but limited in significance. In the USA, this might occur when a state transitions to a new suite of tests which cover the same subject areas as old ones did, e.g. maths and reading, but which include slightly different content elements. Imagine, for instance, that the blueprint for a maths test that was just about to be replaced included calculus but not statistics, whilst the blueprint for the new maths test included statistics but not calculus. Particularly when such tests are used for accountability purposes, teachers will naturally **realign** their teaching: no longer teaching the no-longer-tested elements; and teaching instead the newly-tested elements. In other words, the subject domain that is taught to students will

change (in part) to reflect the (partial) change in the subject domain that is assessed. Thus, over time, following transition to a new suite of tests, students in successive cohorts may genuinely achieve higher levels of attainment *in the new content elements* (and therefore in the new tests overall) as their teachers become increasingly experienced in teaching those new content elements.

Yet another alternative, or complementary, interpretation is that gains on such tests are due to a more manipulative form of realignment, which is so limited in significance as to render those gains entirely inauthentic. The proposition underlying this explanation is that, having realigned their teaching to reflect changes in the assessed subject domain, teachers then develop a deeper appreciation of dimensions of the assessed subject domain that do not *actually* get sampled by the assessment (e.g. aspects of the statistics sub-domain that are not actually assessed). They consequently **reallocate** their instructional resources away from those non-sampled dimensions. With more time and effort devoted to teaching and learning of the sampled dimensions, students in successive cohorts may genuinely achieve higher levels of attainment *on those sampled dimensions*; but only by achieving correspondingly lower levels of attainment on the non-sampled dimensions. Test scores will rise, reflecting higher levels of attainment on the sampled/tested dimensions. But it would be wrong to interpret this as an attainment gain, because any (revealed) gain would be cancelled by a corresponding (hidden) loss, meaning that attainment (overall) would remain the same.

These three concepts – realignment, reallocation, and coaching – are central to understanding the Sawtooth Effect, as well as related effects. They are used, here, in essentially the same way as Koretz uses them (e.g. Koretz, 2008, pp.251-259). However, as described by Koretz, they shade into each other a little. To minimise this shading, and to tailor these concepts to the UK context, the present paper draws these distinctions slightly more sharply than he does. In particular, realignment is treated in the context of syllabus reform as an inevitable and appropriate response to a change in syllabus content, i.e. to a change in the definition of the domain. Consequently, we should *expect* teaching and learning to be realigned towards any new content and away from any omitted content. In contrast, coaching and reallocation are treated as neither inevitable nor entirely appropriate. Even though teachers might engage in such practices without bad intent, they are not constructive practices from an educational perspective, because they focus upon inflating results rather than upon achieving robust understanding/valuable learning. (See below, in the *Mechanisms* section, for further details on realignment, coaching, and reallocation; as well as the formalisation of an additional category, adeptness.)[6]

[6] Koretz (e.g. 2008, p.251) identified a fourth category of questionable techniques – cheating, e.g. revealing test questions to students in advance of the test, or changing their answers following the test. Although an increase in the prevalence of cheating over time certainly could corrupt test scores

UK Analogue

Alastair Pollitt (1998) seems to have been the first to have discussed similar effects occurring in the UK. These were in the context of a challenge that had been faced when a new maths syllabus had been introduced (in 1986) alongside the old version of that syllabus, which was also running in parallel (albeit for the last time). He proposed that:

> When the committee met to recommend boundary marks for the new syllabus there was no 'last year' available for comparison, except the old syllabus. They were aware that some of the content was unfamiliar, not well covered by textbooks and so likely to be rather difficult. What should they have done? If they demanded a level of performance equivalent to that on the old syllabus, ignoring the extra difficulty on the new one, they would be unfairly penalising the schools who had made the transition and unfairly rewarding the ones who had not. I suggest that they recognised this and quite properly 'made an allowance' for the extra difficulty, accepting a lower level of performance for an A or a B grade.

But, what about the following year?

> In 1987, the committee met again. This time there was no old syllabus to worry about, since everyone was on the new one. This time, I suggest, they 'forgot' that a special allowance for unfamiliarity had been made last year and set the 1987 performance standard equal to the lowered 1986 one. Since then year by year comparisons have ensured that the standard today is still that set by special allowance in 1986. We might call this hypothesis 'stepwise standards'.

It is worth emphasising that the pattern described here is not a sawtooth (dropping down then rising up) but a step (simply dropping down). There would seem to be two related reasons for this difference. First, Pollitt's analysis focused on the transition year itself, and the decision to lower the standard in that year. Second, Pollitt was explaining how the transition impact was *managed*, i.e. how potentially problematic consequences were actively mitigated, by dropping the standard in a step-change

---

in a similar way to an increase in the prevalence of coaching (and should therefore not be ignored) it seems less central to our discussion of the Sawtooth Effect. The present report also pays less attention to other problematic practices associated with test-based accountability systems, including: committing far more teaching and learning resources to students likely to score just below a reporting threshold (e.g. grade C in a system that focuses on the percentage of students who achieve grade A* to C, similar to practices associated with the old GCSE grading scale). Although these are important concerns (see Koretz, 2017, pp.69-71), they are not directly related to the Sawtooth Effect, as it is defined in the present report. Having said that, they are relevant to the more general issue of Grade Inflation, and we will return to them in the Postscript.

manner. In other words, the pattern refers to the *exam standard*, rather than to candidates' *performances* or *results* (either immediately, or over time).

This contrasts with how the Sawtooth Effect tends to be discussed in the USA, where the drop is literally *measured*, e.g. in terms of mean percentile ranking against nationally standardised norms. Conversely, for UK exams, the drop is not measured, because the transition impact is anticipated and accommodated, by setting lower grade boundaries. This ensures that candidates in the first year of a new syllabus do not end up with lower grades (i.e. lower than they would have achieved had they followed the old syllabus) *purely as a consequence of being in the 'inaugural' cohort*. This principle for managing standards across transition periods – which has been described as adopting the **Comparable Outcomes** perspective (see Cresswell, 2003) – is now generally accepted (in the UK) as a principle of best practice.[7]

Recent usage in the UK

Although the Sawtooth Effect is an established phenomenon of the educational assessment literature, it has not been extensively researched, even in the USA. In the UK, the effect has featured even less prominently. Indeed, the term itself has only recently been 'imported' to England, via Ofqual's research (Cuff, 2016). Having said that, Ofqual's research has been widely disseminated and discussed, and the Sawtooth Effect has even begun to work its way into public discourse.

Reference to the Sawtooth Effect has also been recorded in Hansard (from the debate *Improving Education Standards*, 29/11/2018)

> *Mike Kane (Wythenshawe and Sale East) (Lab)*: The Minister tells us that success and attainment in the primary school curriculum have gone up, but let us deconstruct that. All the international evidence produced over the past 30 years shows that interventions in the curriculum—and the Minister has had a few—and testing produce disruption to teaching and learning whereby results initially start low, rapidly improve as teachers and students learn what they need to do in order to do well in the tests, then tail off and plateau as this artificial improvement stops. This is known as teaching to the test. He can produce the statistics, but even Ofqual has recognised this problem as the "sawtooth effect". That is what happens when we change the curriculum.

Although this quotation from Hansard provides an example of the Sawtooth Effect being introduced to a debate in order to question the legitimacy of attainment gains over time (on national curriculum tests) – which is in line with usage in the USA –

[7] 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. This can be called the *Comparable Outcomes* perspective and seems unexceptionable." (Cresswell, 2003, p.14)

recent usage in the UK has tended to focus more on the need to manage assessment outcomes across transition periods, by applying the Comparable Outcomes principle. The following passage, from an information piece by Qualifications Wales entitled *Spotlight on Comparable Outcomes* (July, 2018), illustrates this usage:

> In practical terms, we require exam boards to make an adjustment to grade boundaries for the sawtooth effect. When grade boundaries are set, evidence will be considered to see if an adjustment is needed to allow for this effect. Where there is evidence to support the presence of the sawtooth effect, grade boundaries are likely to be lowered, to compensate for the drop in performance of those students sitting new qualifications.

Unpacking the effect

To summarise the story so far: as the term Sawtooth Effect has typically been used in the UK over the past few years, it relates to the issue of standards over time, for large-scale educational assessments that are taken by successive cohorts of candidates, classically from one year to the next. More specifically, it relates to the challenge of managing assessment standards across a period of transition.

Transition

The kind of transition of relevance to the Sawtooth Effect will normally involve a significant change to the syllabus that the assessment is aligned to. However, it may also involve significant change to the approach that is taken to assessment; and it might also involve significant change to the wider curriculum, within which the teaching of the syllabus occurs. The following definitions clarify the intended meaning of these terms in relation to the present analysis:

- **Syllabus transition** – means a change in the nature (i.e. definition) of the domain that is to be assessed. In England, GCSE and A level syllabuses have tended to be 'reformed' every 5 to 10 years, updating subject content and intended learning outcomes, to ensure their continued relevance.[8]
- **Assessment transition** – means a change in the approach that is adopted to assessing a syllabus. Although GCSE and A level exams routinely pose different questions from year to year, successive versions of an exam are still built according to a common blueprint (i.e. using the same question formats and the same question paper structure) and the wider assessment model also remains the same (e.g. with exams lasting the same amount of time, and scheduled at the same time of the year). However, when syllabuses are reformed, assessment approaches may also change. These changes can be

[8] Qualification syllabuses are conveyed through qualification 'specification' documents, and tend nowadays to be referred to as 'specifications'.

radical at times, e.g. if coursework were to be removed, or if a linear structure were to be replaced by a modular one.

- **Curriculum transition** – means a change in the relationship between the syllabus and the wider curriculum. This might involve a change in expectation of how much curriculum time ought to be committed to teaching/learning a particular syllabus; for instance, if a change in performance table metrics were to encourage learners to study fewer subjects. Or it might involve a change in curriculum arrangements during earlier phases; for instance, if all students in England were required to study philosophy in key stage 4, then this would have an anticipatable impact on standards in A level philosophy some years later.[9]

Core features

To warrant talk of the Sawtooth Effect, we would also expect any transition to involve: an **anticipatable** change in the quality of performance of adjacent cohorts; the significance of which is somehow **contestable**, given the presumed nature and cause of that change.

Anticipatable change in performance

It seems to be most helpful to think of the Sawtooth Effect, first and foremost, as an effect upon the quality of *performance* of one cohort of candidates, relative to the quality of *performance* of an adjacent cohort. As noted above, we presume that this effect is attributable to syllabus change, assessment change, and/or curriculum change.[10]

By invoking the Sawtooth Effect, we typically *infer* a change in the quality of performance, from one cohort to the next, rather than necessarily *observing* it directly. In other words, we treat it is as an anticipatable effect; an effect that we may presume to occur, even when there may be limited (or no) direct empirical evidence of its occurrence.

One of the reasons why it may not be possible to directly *observe* a change in the quality of performance, across a period of transition, is that this may be obscured by the nature of the change itself. Imagine, for instance, that an A level chemistry syllabus changed by omitting 50 guided-learning-hours-worth of old content, and by

[9] Note that the syllabus to which a National Curriculum Assessment is aligned is the content of the National Curriculum for the relevant subject and key stage; which is a different sense of 'curriculum' to that described above.

[10] It seems most helpful to think of this as an effect on performances (cf. results, for instance) because, in England, the effect is typically anticipated and accommodated; specifically to ensure that there is no undue impact on results. Having said that, to say that this is an effect on performances is not entirely unproblematic; because, to the extent that a syllabus has changed, we may not be comparing the same *kinds* of performances (e.g. if calculus were to be replaced by statistics).

adding another 50 guided-learning-hours-worth of new content. Clearly, there is no sense in which we could directly observe any change in performance on the omitted content, or on the added content, across the old and new exams (because the omitted content would not be assessed on the new exam, and new content would not have been assessed on the old). Moreover, all other things being equal, we would not expect performance on the content that remained the same to change.[11]

Despite our inability to *observe* change in the quality of performance from one cohort to the next, that does not prevent us from being able to *infer* it. For instance, all other things being equal, we would have no reason to assume that comparable A level chemistry cohorts[12] from adjacent years would perform differently (on average) on the content that remained the same. However, we might well have reason to assume that they would perform differently (on average) on the content that changed. More specifically, we might expect the later cohort to have found it more challenging to master the new content, given its unfamiliarity to them and to their teachers, and we might well expect this handicap to follow through into poorer exam performances, relatively speaking. This kind of inferential reasoning is at the heart of the Sawtooth Effect.

Contestable significance

Talk of the Sawtooth Effect only comes into play when we have some reason to contest the significance of the change in performance; either casting doubt upon its **authenticity**, i.e. whether it reflects a corresponding change in attainment; or, assuming that it does reflect a corresponding change in attainment, whether this represents an **important** change. Our *interpretation* of the change in performance is therefore critical to the Sawtooth Effect. The following examples help to illustrate this.

An inauthentic change in quality of performance, across adjacent cohorts, would be one that related purely to performance on the assessment, without being attributable to any corresponding change in level of attainment. This can occur as a result of coaching, whereby teachers gradually become better at identifying **hacks** or **strategies** for performing successfully on the kinds of tasks that predictably occur on a test. Consequently, over time, candidates become better at scoring marks, but without any corresponding improvement in their understanding of the content

[11] Admittedly, if quality of performance on the content that remained the same *were* to change, then there would be a sense in which we could potentially observe that change directly, by scrutinising exam performances. However, that would still not provide a strong warrant for concluding that the quality of performance overall had changed – i.e. across the entire syllabus – which would involve generalising to quality of performance on the changed content, without any empirical basis for doing so.

[12] Let's say that mean GCSE result scores were equivalent across the two adjacent A level chemistry cohorts, from which we concluded that they were similarly able cohorts.

domain. In a sense, the *performance gain* is real, but we would not want to call it authentic, as it represents *nothing more* than performance gain, devoid of any corresponding gain in robust understanding/valuable learning. Koretz (2008) describes coaching as focusing instruction on small, substantively unimportant details of tasks within an assessment. This would include hacks for identifying correct responses without any substantive understanding at all, e.g. if it were pointed out to students that answers to a certain kind of multiple-choice question can be determined purely by a process of elimination. It would also include strategies for obtaining marks with little or no substantive understanding, e.g. if students were trained to respond to a certain kind of task by regurgitating model answers.

Change in quality of performance, across adjacent cohorts, can also arise as a result of **question spotting**, which inflates results inauthentically when it is combined with reallocation. In this context, teachers gradually become better at identifying assessment content or format sampling patterns, and then tailor their instruction and assessment preparation to reflect these patterns. Consequently, once again, candidates become better over time at scoring marks, but without any corresponding improvement in their understanding of the content domain. A recent study by Ofqual has investigated this phenomenon, in the context of GCSE, AS, and A level exams; identifying the sources of information that teachers consider when predicting exam questions (Holmes, et al, 2020; see Box 1). Baird, et al (2014a, Table 2; 2014b; Table 1) explored similar phenomena, identifying a range of features that might be associated with predictability or unpredictability, and exploring their possible impacts.

A more authentic, yet still contestable, change in performance over time is likely to occur following a change of syllabus content; for example, if a new maths syllabus were to replace calculus with statistics. The first year that candidates are examined on this new syllabus, there will have been no ‘past papers’ to illustrate the approach to examining the new content elements; and there may have been only limited (if any) sample assessment materials to guide teachers and candidates in this respect. Likewise, if teachers had not had to teach statistics to previous cohorts, then they would not have been 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. To the extent that any or all of these possibilities were to become a reality, the first cohort would be disadvantaged in learning the new content elements; a disadvantage that would become gradually smaller over time for successive cohorts. Consequently, all other things being equal, we would anticipate an *authentic* attainment gain, over time, corresponding to better teaching and learning of statistics. However, that gain would still be *contestable*, having occurred from a comparatively lower baseline; lower, that is, relative to the quality of teaching and learning of calculus, for

candidates in the last year of studying the old syllabus. An authentic change – yes (albeit related to only a part of the syllabus). An important change – no.[13]

The key issue, here, is the interpretation of any change in attainment and/or performance over time. Following a substantial change in syllabus content, we might anticipate that the quality of performance of successive cohorts of candidates would improve gradually over time; specifically in relation to the new syllabus content. We might even be prepared to interpret this as an authentic change in level of attainment over time, attributable to better teaching and learning *of the new syllabus content elements*. However, we would not be at liberty to interpret rising results as evidence of better teaching and learning, *per se*. That would be to neglect the fact that those gains were restricted to the new syllabus content, and the fact that they had arisen from a comparatively lower baseline.

Of course, the fact that there may be gains of *contestable* significance during the early years of a reformed syllabus does not rule out the possibility that there may also be *non-contestable* gains, operating simultaneously. Purely for the sake of illustration, let's imagine that a reformed syllabus had been rolled-out at the same time as an initiative designed to improve quality of teaching in that syllabus area, and that this initiative had been successful. In this instance, then, the improvement in pedagogical technique had led to an increase in attainment, for the first cohort of candidates. However, we would also expect there to be a Sawtooth Effect, which would lead to a decrease in attainment for that cohort. Indeed, these two effects might ultimately cancel each other out; the gain due to an increase in teaching quality cancelling out the loss due to a decrease in syllabus/assessment familiarity. This example helps us to appreciate that the Sawtooth Effect is essentially a 'thinking tool' and is not *necessarily* observable.

Finally, it is important to note that a key use of the Sawtooth Effect – as a conceptual thinking tool – is to help us to debate the significance of impacts from syllabus, assessment, and/or curriculum transitions; that is, to help us to decide whether those impacts *ought* to be contested (and managed). Impacts from curriculum change are particularly complicated to evaluate in this sense. Imagine, for instance, that a change in performance table metrics had achieved its intended impact of encouraging learners to study a greater number of subjects; leading to a de facto curriculum reform, in which the vast majority of learners studied around 6 subjects (post-reform) as opposed to around 4 (pre-reform). Clearly, the amount of curriculum time available for studying each subject would be reduced; and, correspondingly, we would anticipate an authentic attainment drop, pre- to post-reform, in each subject

[13] For the sake of simplicity, this analysis glosses over the question of whether it might somehow be 'intrinsically' harder to master the old calculus elements, relative to the new statistics elements (or vice versa).

area. The question for debate would then be whether that authentic attainment drop ought to be contested, and managed, in order to prevent it from impacting negatively upon candidates' results.[14]

[14] We might (perhaps) choose to contest and manage the effect in an attempt to be 'fair' to candidates from adjacent cohorts, who would potentially be in competition for the same higher-level courses (e.g. where pre-reform candidates deferred the application process for a year). That is, we might apply the Comparable Outcomes principle to boost post-reform candidates' subject grades, to ensure that they were not penalised unfairly for having been in the 'inaugural' cohort.

Box 1: Factors influencing teacher predictions (Holmes, et al, 2020)

Armed with their own intuitions, and with exam materials from the previous four years, teachers of GCSE history, AS government & politics, and A level psychology – approximately 10 per subject – were invited to make independent predictions of the questions that they thought might appear in subsequent years, and to identify the factors that influenced their predictions. Twenty five factors were identified, in total, grouped within five clusters (factors 4, 5, 6, and 8 tended to be mentioned more frequently than others, across the three subjects):

A. Factors related to appearance of questions/topics on past papers

1. High frequency of topic appearance makes it likely (to come up)
2. High frequency of question type appearance makes it likely (to come up)
3. Topic/question type has come up frequently, but in a different form/place
4. Past patterns of question type/topic cycling lead to this topic
5. Non-appearance last year/recent years/ever increases chance of appearance

B. Factors related to the content in the specification document

6. Importance/centrality in specification makes it likely it will come up
7. Alignment of wording to specification content

C. Factors related to the appropriateness of topic for the type of question

8. Topic/question type fits position on paper (size, type – event/treaty etc)
9. Topic difficulty fits position on paper or question tariff
10. Differentiating between candidates of different abilities

D. Factors related to the logic of whole papers

11. Need to cover part of content not assessed elsewhere/content balance
12. Question effective at assessing a substantial portion of syllabus
13. Need to balance types of questions in unstructured sections of papers
14. Logical/chronological order of questions

E. Factors related to the age of the syllabus

15. Long-lived specification and likelihood of unusual or random questions
16. New specification with limited past papers means existing questions will not be used
17. Avoidance of new topics in early days of specification
18. New topics on specification need to be assessed in the early live papers

F. Factors revolving around other resources such as textbooks, sample assessments, availability of sources or topicality

19. Structure of textbook
20. Topics appearing in textbooks make these topics more likely
21. Language used in textbooks
22. Exclusion of example questions in textbooks/Sample Assessment Materials

23. Use of example questions in textbooks/Sample Assessment Materials
24. Availability of source material for this topic
25. Topicality of question (given timelines for paper production)

This curriculum-related example raises questions that behove us to unpack the Sawtooth Effect in greater depth, and to consider how narrowly or broadly we are prepared to apply the term. There are two issues here. First, it should be evident that the impact of curriculum change is not necessarily negative. If more curriculum time were to be made available for studying each syllabus, then we would anticipate an authentic attainment gain, pre-reform to post-reform, in each subject area. This would imply a contestable *rise* in performance, across the period from the last year pre-reform to the first year post-reform, not a contestable *drop*. If so, then would we still want to invoke the Sawtooth Effect under these circumstances? Second, it should also be evident that the impact of curriculum change would remain the same from the first year post-reform to successive years. In other words, we would not anticipate a contestable change in performance following the transition year. Again, if so, then would we still want to invoke the Sawtooth Effect under these circumstances?

Separable effects

In order to unpack the Sawtooth Effect, we need to recognise that it actually comprises two distinct effects: a **Disruption Effect**; and an **Enhancement Effect**.

The Disruption Effect corresponds to the change in quality of performance that is presumed, or observed, to occur across the transition from the last administration of the old arrangements to the first administration of the new ones. It is essentially discrete, and might be substantial in magnitude. We would expect this change to be negative, i.e. to reflect a sudden drop in performance.

The Enhancement Effect corresponds to the change in quality of performance that is presumed, or observed, to occur from the first administration of the new arrangements to successive administrations. This effect is essentially continuous, and is less likely to be substantial even from year 1 to year 2. We would expect this change to be positive, i.e. to reflect a gradual rise in performance.

Mechanisms

If Sawtooth Effects are concerned with impacts on performances that are somehow contestable, given the presumed nature and causes of those impacts, then this behoves us to reflect more deeply upon the mechanisms by which such effects might occur. Table 1 presents a summary of the kinds of impacts that tend to be implicated in discussions concerning the Sawtooth Effect, broken down according to whether they are presumed to influence the Disruption Effect, the Enhancement Effect, or both.

The first substantive column of Table 1 is headed Nature (and Causes) of Presumed Impacts. It identifies various different kinds of impact, providing an example of a

possible cause of each kind of impact in parenthesis. These impacts fall into one of three main categories:

i. authentic impacts on teaching and learning (and attainment);
ii. inauthentic impacts on performance (alone); and
iii. authentic impacts on baseline attainment.

When one of the three transitions (syllabus, assessment, or curriculum) impacts upon teaching and learning, the implication is that this will impact upon attainment (and therefore also upon performance in the assessment). It is assumed that this impact on attainment is authentic, although its significance will be at least somewhat contestable, for one reason or another. The third category of impacts that feature within Table 1 recognises the (admittedly unusual) situation of impacts affecting the baseline level of attainment of a cohort of students, before they even begin to study a syllabus.

Perhaps the most important observation to make from Table 1 is that only 3 of these 6 presumed impacts (the 1st, 2nd, and 4th) trace out the classic sawtooth pattern; that is, a sudden performance drop followed by a gradual performance rise. They are all essentially familiarity effects. The remaining 3 presumed impacts are also part of the broader conceptual territory of the Sawtooth Effect. That is, they also refer to anticipatable changes in the performance of adjacent cohorts, which occur across periods of transition, and which are of contestable significance. But they would seem to be restricted in influence to the Disruption Effect, and the changes in question are not necessarily in a negative direction. In other words, they reside in the territory of the effect despite not embodying the Sawtooth Effect phenomenon in its entirety.[15] We might call this the territory of **contestable transition impacts** within which the Sawtooth Effect is its archetypal case. It is important to recognise that these presumed impacts all operate in a common conceptual territory because of how they will inevitably interact. Being able to identify and anticipate such interactions is critical to being able to manage transition impacts convincingly.

[15] The tooth, but not the whole tooth.

	Nature (and Causes) of Presumed Impacts	**Anticipated Disruption Effect** in the first year following the transition	**Anticipated Enhancement Effect** in subsequent years following the transition
Change in syllabus content	**1. Authentic impact on teaching and learning** of new content elements (e.g. due to content unfamiliarity, and lack of availability of teaching resources).	**Attainment Drop**: New content elements are likely to be taught less effectively, and learned less effectively, than old (omitted/retained) content areas.	**Attainment Rise**: Over time, new content elements will come to be taught, and learned, as effectively as old (omitted/retained) content areas. [realignment]
	2. Inauthentic impact on performance in assessments of new content elements (e.g. due to coaching hacks and strategies, and/or to teaching/learning reallocation).	**Performance Drop**: Content-specific assessment hacks and strategies unlikely to have been developed for new content elements; and assessment content sampling patterns will not (yet) be predictable.	**Performance Rise**: Over time, assessment hacks and strategies may come to be developed for new content elements. [coaching] Improvements may follow from the identification of assessment content sampling patterns. [reallocation]
Change in assessment approach	**3. Authentic impact on teaching and learning** across all syllabus content (e.g. with change in assessment structure from linear to modular).	**Attainment Change**: New structure may have backwash impact on teaching and learning. Note that this could be either negative (Disruption) *or* positive (Enhancement).	[Possible incremental change in magnitude of this impact over time as the change becomes embedded.]

	Nature (and Causes) of Presumed Impacts	**Anticipated Disruption Effect** in the first year following the transition	**Anticipated Enhancement Effect** in subsequent years following the transition
	4. Inauthentic impact on performance in assessments across all syllabus content (e.g. with change in assessment format from mainly multiple-choice to mainly short-answer).	**Performance Drop**: New (unfamiliar) assessment structures/formats are likely to be tackled less effectively than established (familiar) ones. Also, format-specific assessment hacks and strategies unlikely to have been developed for new assessment formats; and assessment format sampling patterns will not (yet) be predictable.	**Performance Rise**: Over time, new assessment structures/formats will come to be tackled as effectively as established ones (becoming just as familiar). [adeptness] Over time, assessment hacks and strategies may come to be developed for new assessment formats. [coaching] Improvements may follow from the identification of assessment format sampling patterns. [reallocation]
Change in curriculum organisation	**5. Authentic impact on teaching and learning** across all syllabus content (e.g. with 15% more/less time available to study syllabus).	**Attainment Change**: More or less time committed to teaching and learning will impact on attainment. Note that this could be either negative (Disruption) *or* positive (Enhancement).	No effect.

	Nature (and Causes) of Presumed Impacts	**Anticipated Disruption Effect** in the first year following the transition	**Anticipated Enhancement Effect** in subsequent years following the transition
	6. Authentic impact on baseline attainment of all students (e.g. with study of subject made compulsory, or withdrawn, at lower phase).	**Attainment Change**: Different (course-entry) baseline attainment levels will carry through (to course exit). Note that this could be either negative (Disruption) *or* positive (Enhancement).	No effect.

Table 1. Mechanisms underlying Sawtooth Effects, and related effects

Highlighted in blue, in the final column of Table 4, are the mechanisms of the Enhancement Effect, i.e. the reasons why we might expect performance/attainment to rise over time, post-transition. All of these are contestable, the first in terms of importance, and the remainder in terms of authenticity:

1. **realignment**, i.e. as new content elements become familiar over time, teachers become better at teaching them, resources improve, and (as a consequence) learners come to learn them better;
2. **adeptness**, i.e. as new assessment structures/assessment formats become familiar over time, teachers enable learners to become increasingly adept at recognising, navigating and responding to these new assessment demands, and thus more effective at demonstrating their actual levels of attainment;[16,17]

[16] This is distinct from the realignment effect because we would expect increasing adeptness to result in candidates becoming better at *demonstrating* their levels of attainment, even if those levels of attainment were to remain constant over time.

[17] Shepard (1997) hinted at a subtly different explanation, which might also be classified under this adeptness category. As new task formats become increasingly familiar to candidates, they become more likely to be able to score marks with nothing but a fragile, fledgling grasp of the concepts at stake; certainly not enough of a grasp to reflect a robust understanding of those concepts, and therefore not enough of a grasp to be genuinely worthy of those marks. In other words, as successive cohorts become increasingly adept with the new task formats, they become increasingly likely to score marks fortuitously; even without having been explicitly coached in hacks or strategies. Once again, we might anticipate that such effects would occur even if level of attainment remained constant over time. (Admittedly, there might be a case for identifying this as a distinct category, in its own right, located somewhere in between adeptness and coaching.)

3. **coaching**, i.e. as new content elements/assessment formats become familiar over time, teachers begin to identify (and instruct their students in) hacks and strategies for scoring marks (without any corresponding increment in attainment);
4. **reallocation**, i.e. as new content elements/assessment formats become familiar over time, teachers become better at question spotting, reallocating their instructional resources towards those elements/formats that are most frequently sampled (and away from those elements/formats that are least frequently sampled).

Closer inspection of these categories reveals that the Enhancement Effect actually comprises two distinct effects: a **Recovery Effect**, and an **Augmentation Effect**. The Recovery Effect is due to realignment and adeptness. Realignment is the process by which teachers (and learners) get back up to scratch in terms of teaching (and learning); and adeptness is the process by which teachers (and learners) get back up to scratch in terms of assessment. In other words, the Recovery Effect enables successive cohorts to recover from the transition to new arrangements. The Augmentation Effect goes one step further. Via coaching and reallocation, it takes successive cohorts beyond recovery, such that they become able to demonstrate performances of higher quality than might otherwise be associated with their levels of attainment. The Augmentation Effect is therefore more malign.

The Recovery Effect will kick in straight away, and we would expect teachers in the second year post-reform to prepare their students significantly more effectively than in the first year post-reform. The Augmentation Effect is likely to be somewhat delayed, as it will take time for dimensions of predictability to become apparent. Figure 3 illustrates the anticipated chronology of the Augmentation Effect relative to the Recovery Effect. As the Recovery Effect begins to decrease, the Augmentation Effect is likely to to increase.

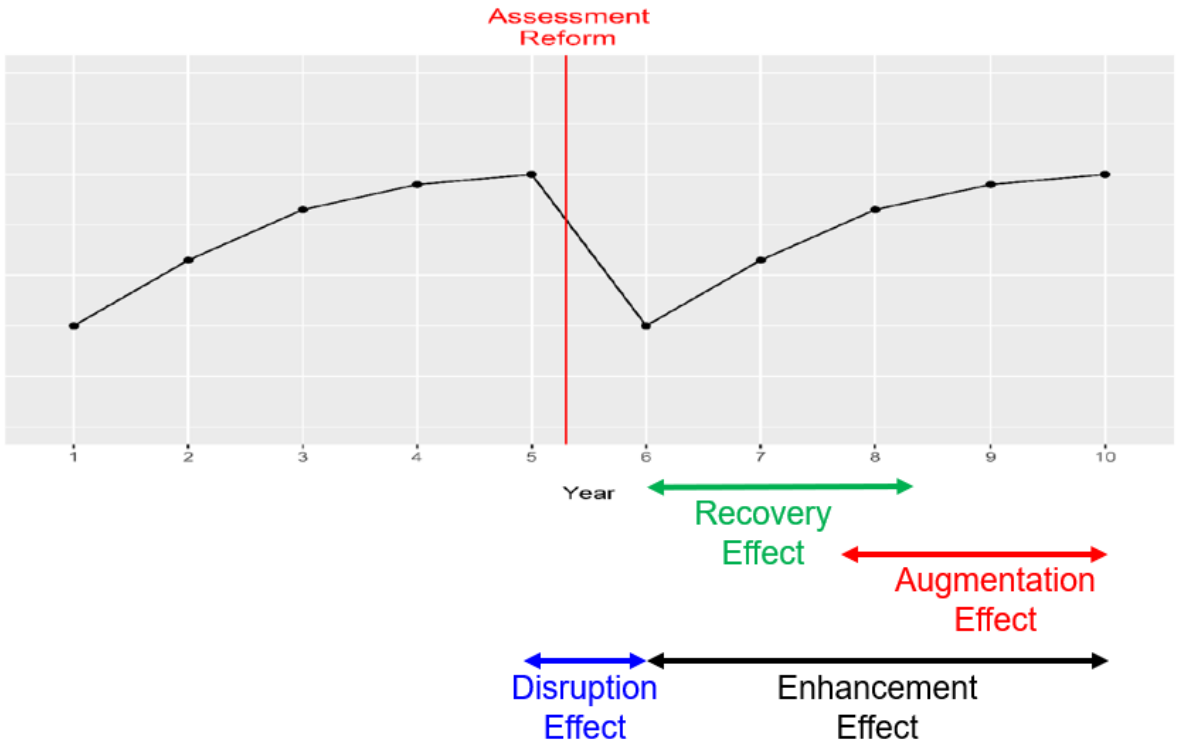

Figure 3. Illustration of anticipated chronology of Sawtooth Sub-Effects

Summary

As it has come to be used in the UK over the last few years, the term Sawtooth Effect relates to the issue of standards over time, for large-scale educational assessments that are taken by successive cohorts of candidates, classically from one year to the next. It refers to an anticipatable change in the quality of performance of adjacent cohorts, which occurs across a period of transition in syllabus, assessment, and/or curriculum arrangements, and which is of contestable significance. It comprises two distinct effects: a discrete performance Disruption Effect, which occurs across the period from the last administration of old arrangements to the first administration of new ones; and a continuous performance Enhancement Effect, which occurs across a more extended period, from the first administration of new arrangements to successive administrations. The Enhancement Effect can also be deconstructed into two distinct effects: a Recovery Effect, which operates via realignment and adeptness; and an Augmentation Effect, which operates via coaching and reallocation.

It is useful to draw a distinction between the Sawtooth Effect phenomenon and the broader landscape of contestable transition impacts. The phenomenon refers to a restricted set of circumstances in which a particular causal factor (or interaction of factors) has both: a Disruption Effect, leading to a sudden performance drop; and an

Enhancement Effect, leading to a gradual performance gain. The Sawtooth Effect phenomenon is the archetypal case in a broader landscape of contestable transition impacts. This territory also includes circumstances that might not necessarily result in an Enhancement Effect (beyond the Disruption Effect); and for which the Disruption Effect might operate in either direction (resulting in either a drop or a rise).

4 The management of transition impacts

As discussed in the USA[18], the Sawtooth Effect tends to be invoked when casting doubt upon the legitimacy of score gains on accountability tests, introducing the alternative hypothesis of **Score Inflation**.[19] Conversely, in the UK, the Sawtooth Effect tends to be invoked when maintaining standards across periods of transition, to recommend applying the Comparable Outcomes principle. This is partly to counter the threat of **Grade Inflation**; but it is primarily to counter the threat of **unfairness**. This is the issue of unfairness to candidates from 'inaugural' cohorts, were they to have received lower results for no other reason than having had to study in the wake of a transition in syllabus, assessment, and/or curriculum arrangements. Or, to put it more starkly, for no other reason than having been born in one academic year, as opposed to the previous one.

In terms of the various categories of fairness identified by Nisbet and Shaw (2019), the Comparable Outcomes principle would appear to be grounded in value judgements that are both **retributive** (in terms of ensuring that candidates are appropriately rewarded) and **relational** (in terms of treating like cases alike). First, the principle would seem to be retributive in the sense of making sure that candidates receive their just deserts for the time and effort that they put into their course of learning; even when that learning may have been handicapped by factors beyond their control, and their overall level of attainment is not as high as it would have been otherwise. Second, the principle would seem to be relational in the sense of ensuring that demographically equivalent cohorts are treated equivalently, by engineering equivalent grade distributions.

The threat of unfairness is considered serious enough to warrant preventative action when grading GCSE and A level exams because of how exam results are used. In both cases, results are used as a basis for selection to higher level courses, and/or to employment, and it is not uncommon for candidates from adjacent cohorts to be in competition for the same opportunities. Conversely, results from national curriculum tests are not used in this way, and the corresponding risk of unfair selection decisions does not arise. This is why the Comparable Outcomes principle is not applied when linking standards on national curriculum tests across periods of transition. Transition impacts still need to be managed, though; albeit primarily via a communications strategy, i.e. by managing the interpretation of results. It is therefore helpful to distinguish between two different approaches to managing transition

[18] For example, by Dan Koretz (2008; 2017).

[19] According to Koretz, Score Inflation "refers to increases in scores that do not signal a commensurate increase in proficiency in the domain of interest" (Koretz, 2008, p.34).

impacts: managing outcomes; and managing interpretations. The following sections will consider each of these in turn.

Managing outcomes

To the extent that it is reasonable to treat transition impacts as though they were universal effects – in the sense of being likely to affect all, or the large majority of, candidates from a particular cohort in a similar way – it will be possible to manage them, during the grade awarding process, by deciding where to locate grade boundary marks. This assumption is critical to how we apply the Comparable Outcomes principle.

Managing the Disruption Effect

The primary rationale underpinning the Comparable Outcomes principle is to be fair to candidates following a syllabus, assessment, and/or curriculum transition. An entry[20] in *The Ofqual blog* from 2017 put it like this:

> When qualifications change, we follow the principle of comparable outcomes – this means that if the national cohort for a subject is similar (in terms of past performance) to last year, then results should also be similar at a national level in that subject. So exam boards will control for the impact of the changes such that this year's cohort is not unfairly disadvantaged. They will be relying heavily on the statistical evidence to do this, but also using senior examiners to check the grade boundaries that the statistics are pointing to.

Applying the Comparable Outcomes principle during the first year following a transition is tantamount to saying: even though the level of attainment of the post-transition cohort might (in some 'fuzzy' sense) actually have fallen, relative to the level of attainment of the pre-transition cohort, we will manage the grade awarding process as though it had remained the same.[21] We say this, having anticipated a drop in attainment, while contesting its significance.

---

[20] *The Ofqual blog: Comparable outcomes and new A levels*. (Cath Jadhav, 10 March, 2017.) Available from: https://ofqual.blog.gov.uk/2017/03/10/comparable-outcomes-and-new-a-levels/

[21] Putting it like this conceals the additional unstated assumption that we have no other reason to expect the two cohorts to differ in level of attainment. Of course, there *might* be other reasons for expecting the two cohorts to differ; for instance, demographic reasons, such as the composition of one cohort tending to be more skewed towards higher-achieving students than the other. Consequently, in practice, this principle is applied only after having controlled for what is considered to be the most significant demographic factor: prior attainment. This is achieved via prediction matrices, which ensure that, on average, candidates with similar prior attainment measures will end up with similar grade outcomes, pre- versus post-transition.

Technically, it is not critical to any application of the Comparable Outcomes principle that there actually *has* been a drop in attainment, pre- to post-transition; it simply accommodates the *possibility* of a drop. In fact, for reasons that are largely theoretical, there could be no way to determine the 'truth' concerning any change in level of attainment across a period of transition.[22] More pragmatically, even if a transition happened to be correlated with a genuine, cohort-wide improvement in teaching quality – one that more than cancelled out any Sawtooth Effect – we would still have no way of knowing this; because we would have no independent instrument for measuring the net effect, let alone for teasing apart its component effects. So, the fairest policy – according to the Comparable Outcomes principle – is simply to manage the grade awarding process, during periods of transition, as though there were no change at all (assuming that incidental demographic changes between cohorts have been controlled for).

Managing the grade awarding process as though there were no change at all in level of attainment, from one cohort to the next, straightforwardly deals with the complication of potential interaction effects between contestable transition impacts. If, for instance, the explicit intention of a change in assessment structure were to have a positive backwash impact on teaching and learning, then we might anticipate that attainment would rise, post-transition; again, potentially more-than-cancelling-out any Sawtooth Effect. Yet, the Comparable Outcomes logic would come into play once more: why should the post-transition cohort end up with higher grades than the pre-transition cohort, for reasons that are purely attributable to the way in which their teaching, learning, and assessment happened to be structured that year?

Of course, applying the Comparable Outcomes principle under such circumstances makes it clear that it would not be possible to evaluate the **education policy** underlying the transition in terms of exam result trends, pre- to post-transition; for the simple reason that any effect of the transition would be factored out of results. Yet, this only makes explicit what would be true anyway, given that there is no way to determine the 'truth' concerning changes in attainment across transition periods.

Managing the Enhancement Effect

If it is fair to factor contestable transition impacts out of results in the year just following a transition, then it is fair to factor them out of results in subsequent years, too. The underlying logic is identical. Moreover, as indicated in Table 1, we have good reason to anticipate that such effects will occur during the early years following

[22] In this sentence, the word 'truth' is expressed in scare-quotes to emphasise that, from a theoretical point of view, the idea of a change in *level* of attainment (pre- versus post-transition) can only be conceptualised in a loose 'fuzzy' sense, because the *nature* of the attainment (construct) has changed significantly (see, for instance, Newton, 1997; Newton, 2010).

a transition; in a direction that counteracts the Disruption Effect. What is less clear is exactly how long such effects will persist (although research by Cuff, 2016, provides important insights). And what is still somewhat unclear is exactly how such effects ought to be managed.

Fairness

The most pressing considerations when managing outcomes following a transition period are the impacts of growing familiarity with the new arrangements: familiarity with new content elements, and how best to teach them; and familiarity with new assessment structures/formats, and how best to approach them. Experience of, and feedback from, each new session will provide teachers with important insights for future exam sessions; which they will pass on to the next cohort of learners.

The need to adjust to new content elements and new assessment structures/formats will be immediately apparent to all concerned. New content elements will be revealed once the reformed syllabus has been disseminated, prior to first teaching; and new assessment structures/formats may also be explained and exemplified at this stage, assuming that sample assessment materials have been provided. Teachers will be strongly motivated to overcome the effects of unfamiliarity as soon as possible. Consequently, we might expect the Recovery Effect – which operates via realignment and adeptness – to be especially significant during the early years of a reformed qualification, tailing off after a certain period of time.

The primary justification for any decision to manage outcomes in order to accommodate anticipated realignment and adeptness effects would relate to fairness: why should the year 2 cohort end up with higher grades than the year 1 cohort, for reasons that are purely attributable to the year 1 cohort having had less opportunity to become familiar with the new arrangements? Instead, for reasons of fairness, we might decide that even though the level of attainment of the year 2 cohort might actually have risen, relative to the level of attainment of the year 1 cohort, we will manage the grade awarding process as though it had remained the same. In other words, we might decide to continue to apply the Comparable Outcomes principle for a few years, until realignment and adeptness effects had tailed off.

As noted earlier, we would have no way to distinguish between these contestable realignment/adeptness effects and any non-contestable effects that happened also to cause attainment to rise or fall simultaneously. We would simply manage the grade awarding process during the early years of a reformed qualification *as though* attainment had remained the same (assuming that incidental demographic changes between cohorts had been controlled for).

Grade inflation

As familiarity with new arrangements increases, the Augmentation Effect may come into play, causing the quality of performances to rise over time, albeit attributable purely to coaching (hacks and strategies) or to reallocation (question spotting). The potential for coaching and reallocation would presumably take longer to become evident; as the predictability of new arrangements became increasingly apparent over time. Consequently, it seems likely that the impact of the Augmentation Effect might *step up* over the early years of a reformed qualification, rather than tail off.

In response to the threat of Grade Inflation attributable to coaching and reallocation, a case might be made for continuing to apply the Comparable Outcomes principle beyond the 'first few years' of a reformed qualification. Of course, the longer the Comparable Outcomes principle is applied, the less possible it becomes to recognise any net, non-contestable, authentic, cohort-level attainment change over time. Continuing to apply the Comparable Outcomes principle over an extended period of time, if not indefinitely, might be to take a strong and defensible stance in response the threat of Grade Inflation, as well as unfairness, but it would be at the expense of being able to recognise attainment trends over time via exam results.[23]

In fact, the first time that it was decided – as a matter of national policy – to continue to apply the Comparable Outcomes principle beyond the first year following a transition, the justification was framed in terms of Grade Inflation rather than fairness:[24]

> The key points are:
>
> In summer 2012 we will continue to prioritise comparable outcomes in A levels, to avoid grade inflation
>
> In new GCSEs awarded for the second time in summer 2012, we will continue to prioritise comparable outcomes, to avoid grade inflation
>
> […] Therefore, to avoid grade inflation in 2012 and beyond, we've agreed with exam boards that they will continue to prioritise comparable outcomes.

And this is true, of course. The approach that we would take to manage outcomes on a principle of fairness is the same as the approach that we would take to manage outcomes in order to avoid Grade Inflation – the same approach, based upon the

---

[23] That is, even *between* transition periods, where this is theoretically more plausible than *across* transition periods. Having said that, even *between* transition periods, the interpretation of result trend lines is far from straightforward (e.g. Newton, 1997; Koretz, 2008).

[24] This quotation comes from an Ofqual document entitled: *GCSEs and A Levels in Summer 2012: Our approach to setting and maintaining standards*. Available from: https://dera.ioe.ac.uk//15397/1/2012-05-09-maintaining-standards-in-summer-2012.pdf

same Comparable Outcomes principle. Perhaps the only difference is that the Grade Inflation characterisation more forcefully begs the question of how long to continue this management process, suggesting that it might be appropriate to do so indefinitely. We will return to this issue in the Postscript.

Precautionary approach

It is important to appreciate that applying the Comparable Outcomes principle is a fairly crude, precautionary tactic, that we adopt in order to address a likely threat of unfairness (or Grade Inflation) while never being in a position to determine the 'true' state of affairs in relation to cohort-level changes in attainment over time. This is why they were described, earlier, as 'anticipatable' (in a general sense) rather than 'predictable' (with any confidence). In other words, although we have strong logical and empirical grounds for anticipating the occurrence of contestable transition impacts, we should be less confident in attempting to predict exactly how (or even whether) they are likely to operate in any particular transition context.

For instance, changes to assessment approaches are sometimes intended specifically to improve the quality of teaching and learning, e.g. a transition from linear to modular assessment, or a transition from multiple-choice testing to performance assessment. If so, then we would predict authentic, cohort-level attainment gains (albeit contestable ones). Unfortunately, it is rarely, if ever, possible to judge definitively whether such changes have actually accomplished their intended impacts. Furthermore, for certain structural changes, such as switching from linear to modular assessments, there is persuasive evidence to suggest that they may not always do so (see Baird, et al, 2019).

Finally, we would be on even shakier ground attempting to predict how a variety of contestable effects might interact with each other, and/or with non-contestable effects. In short, any decision to manage outcomes, by applying the Comparable Outcomes principle, is taken from an unavoidable position of nescience. In the absence of any dependable knowledge concerning the occurrence of a net, non-contestable, authentic, cohort-level attainment gain or loss, we simply manage outcomes as though there had been neither (assuming that incidental demographic changes between cohorts have been controlled for).

Challenges

Earlier, we noted an important assumption when applying the Comparable Outcomes principle – that the transition impacts in question are **universal** effects, affecting all, or the large majority of, candidates from a particular cohort in a similar

way. This is because the management process – the location of grade boundary marks – applies in the same way to all candidates within an exam cohort. [25]

If transition impacts operate differentially, rather than universally, then it becomes much harder (if not impossible) to manage those impacts adequately. An example might be useful here. Imagine that a change in exam-paper-writing personnel for GCSE maths happened to make an exam paper 2 marks easier for boys, but not for girls, owing to a variety of subtle choices concerning the contexts within which questions were set, which no-one noticed. Without drilling down into the data, it might look as though the paper had been 1 mark easier, on average, for all candidates in the cohort; and it might be decided to set grade boundary marks 1 mark higher than in the previous year. Yet, clearly, this would not resolve the issue. In a sense, it would mean that boys would still have a 1-mark bonus, whilst girls would still have a 1-mark penalty. But a more appropriate way of thinking about this issue is that there is simply no way to resolve it adequately when locating grade boundary marks, because the bias has corrupted the rank ordering of candidates. Exactly the same issue would arise if transition impacts were to affect subgroups of the cohort differentially.

Fortunately, we have some reason to anticipate that realignment and adeptness effects are likely to be fairly universal. When syllabus, assessment, and/or curriculum arrangements change, they tend to change for everyone concerned. If so, then everyone is likely to be affected by them in a similar way, and everyone will be motivated to adjust to those new circumstances as rapidly as possible. For example, where a syllabus drops calculus and replaces it with statistics, we can safely predict that all teachers will be motivated to get up to speed with teaching statistics. True, it might take some teachers far less time to get up to speed than others. But, at least we can expect there to be a general direction of travel, in which all teachers would be moving.

Differential effects

There is an important complication, however, and this relates to the phenomenon of **cohort churn**. In England, a small number of awarding bodies compete for market

[25] This assumption is not unique to applying the Comparable Outcomes principle; it is an assumption that underpins grade awarding, per se. During grade awarding, where all candidates sit exactly the same exam paper, we might assume that any impact (of differential difficulty) would be universal for all candidates; or, at least, universal for all candidates at any particular mark point (this allows us to make slightly different adjustments for candidates at different grade boundaries). However, even in this situation, the assumption of universality is unlikely to be watertight. In other words, even if the large majority of borderline grade 7 candidates were to find a particular exam paper abnormally challenging, there might still be a small minority of candidates who did not. In this situation, utilitarian considerations would recommend setting a lower grade 7 boundary, to be fair to the largest number of candidates; even though this would mean that some candidates would benefit unduly.

share, offering exchangeable (albeit significantly different) versions of the same qualifications. Often, when syllabuses are reformed, cohorts remain fairly stable, i.e. schools decide to remain with the same awarding body, pre- and post-reform. Sometimes, though, the post-reform syllabus of a particular awarding body will seem particularly attractive, and many more schools will opt for it. In this circumstance, it is possible that the degree of syllabus change will be greater for the incoming schools than for the remaining schools. If so, then the realignment challenge might be greater, and the anticipated Sawtooth Effect more pronounced, for those incoming schools. The larger the difference in size of effect, and the more similar the ratio of incoming to remaining schools, the more controversial any particular Comparable Outcomes adjustment might be.

Turning from realignment/adeptness to coaching and reallocation, the situation becomes more problematic. Indeed, it seems quite likely that coaching and reallocation may operate quite differently across schools within an exam cohort. As noted earlier, these effects are somewhat malign, particularly the coaching effect, and teachers are likely to differ in their willingness to engage in them. If only certain teachers/schools were to seek to inflate their students' marks via coaching or reallocation, then any attempt to counter such effects, i.e. by raising grade boundaries, would be inadequate. In effect, it would penalise those teachers/schools who did not seek to inflate their students' marks in this manner. The bottom line is that grade boundary adjustment cannot effectively be used to rectify bias in exam results that affects only a proportion of the cohort.

Finally, it is worth noting the possibility of a **staggered** Sawtooth Effect. This could occur if cohort churn were not predominantly restricted to the first year of a new syllabus, but occurred gradually, with a substantial influx of new schools each year. By way of example, the realignment effect might well have trailed off by year 4, for schools that offered the reformed syllabus from its outset. Yet, the same effect would be at a maximum in year 4, for schools that had only just switched to the syllabus. Again, this is just another case of differential bias – bias that affects only a proportion of the cohort – and no grade boundary adjustment could effectively rectify it.

These complicating effects of cohort churn and staggering are likely to be exaggerated when syllabus change is compounded by curriculum change; for instance, if there happened to be a change in the A level economics syllabus, combined with A level business studies and BTEC business studies being withdrawn. Under such circumstances, we can imagine an influx of business studies teachers/candidates to the new economics syllabus; but it would be very hard indeed to predict the impact of Sawtooth Effects for these very different subgroups of the population, for whom differential impacts would be expected.

Fixed criteria

Is it legitimate to attempt to manage outcomes, by adopting the Comparable Outcomes perspective, when assessment standards are explicitly specified in terms of fixed criteria? This challenge is particularly salient in relation to the Enhancement Effect.[26] It is most problematic when standards are specified in line with the competence-based assessment tradition of writing standards in terms of assessment criteria nested within learning outcomes nested within units. This is most problematic because any allowance for growing familiarity would have to be made at the level of individual learning outcomes, if not individual assessment criteria. That is, teachers would need to make these judgements; and in the same way, across all learners, across all schools. In practice, it is hard to see how this could be expected to work.

Taking one step back, however, there is an important prior question to ask concerning whether increases in attainment during the early years of a new 'fixed criteria' syllabus – which are purely due to growing familiarity – ought to be considered of contestable significance in the same way (or to the same extent) as they might be for GCSE and A level exams. If not, then the argument for applying a Comparable Outcomes adjustment would not hold anyhow. In the GCSE and A level case, the issue turns principally on fairness to candidates from adjacent cohorts who might subsequently be in competition for the same opportunities. The unstated rationale, here, is that even an authentic change in attainment level – due to either the Disruption Effect or the Enhancement Effect – would not mean that candidates from one cohort were significantly more or less deserving of those subsequent opportunities than any other, nor significantly more or less likely to be able to exploit those opportunities successfully.

For qualifications that are strongly competence based, in the sense of certifying competence-to-practise in an occupational role, it might be hard to mount a similar case. In this context, it might seem that the overriding concern is whether a learner has reached a critical threshold of competence, *regardless* of the challenges they might have faced, or the support that they might have received, in reaching that level of competence.[27] Even if a fairness-based argument in favour of applying the Comparable Outcomes principle were to be constructed – which, arguably, it could be – it is harder to imagine it being capable of winning a similar level of credibility with stakeholders and members of the public.

---

[26] This is because, when standards are defined in terms of fixed criteria, the specification of new criteria (for a new syllabus) is more likely to be treated having set a *new* standard, intentionally distinct from the old one.

[27] I use the term 'competence' here, as it is more familiar in this context. However, I use it to mean the same as 'attainment' (with the implication that both are distinct from 'performance', which merely provides *evidence* of competence/attainment).

For qualifications whose standards are specified in line with the competence-based assessment tradition – but that do not certify competence-to-practise, and that typically support progression into higher-level courses – the case is less clear.[28] For such qualifications, the likelihood of being able to develop a credible fairness-based argument in favour of applying the Comparable Outcomes principle is potentially far higher. Yet, it might still be far from obvious how to apply the principle, in practice.

Similar issues were faced during the 1990s, for GCSEs and A levels with coursework. Coursework components applied exactly the same criteria for the award of marks from one year to the next, so it seemed only natural not to change their grade boundaries. Yet, performance on those components rose radically over time, far more dramatically than for exam components, and for reasons that in retrospect we can feel fairly confident in attributing largely to increasing adeptness, reallocation, and coaching (i.e. to contestable impacts). Where coursework boundaries were held constant over time, the only way to manage this effect was to raise standards on exam components, to balance this out. This proved not to be entirely satisfactory.

Managing interpretations

As just discussed, it may not always be considered appropriate (or possible) to apply the Comparable Outcomes principle; even when there might be good reason to anticipate a Sawtooth Effect. If so, then the effect may need to be managed differently; most obviously by managing the interpretation of assessment outcomes.

Managing the Disruption Effect

When new national curriculum tests were introduced, to assess the new national curriculum for mathematics and reading, the idea of managing the maintenance of standards by applying the Comparable Outcomes principle was judged to be inappropriate. This was because there had been an explicit intention to set new, explicitly higher standards on these tests. The Department for Education (2016) made this clear in its release of both provisional and final statistics covering the 2016 assessments, for example:

> Children sitting key stage 2 tests in 2016 were the first to be taught and assessed under the new national curriculum. The expected standard has been raised and the accountability framework for schools has also changed. These changes mean that the expected standard this year is higher and not comparable with the expected standard used in previous year's statistics. It

[28] Especially qualifications that are taken alongside GCSEs and A levels, and that might even support progression to exactly the same higher-level courses.

> would therefore be incorrect and misleading to make direct comparisons showing changes over time.

Whether or not assessment outcomes have been managed to accommodate the Disruption Effect, if a transition has occurred that has disrupted the interpretation of results from one cohort to the next, then it is clearly good practice to attempt to manage interpretations, via some kind of communications strategy. This raises an important question concerning where responsibility for managing interpretations ought to lie; particularly where a variety of players might be implicated (i.e. assessment agencies, regulators, producers of aggregated statistics).

Managing the Enhancement Effect

Since there was to be no formal linking of old and new trend lines, following the transition to new national curriculum testing arrangements in 2016, the idea that a sawtooth pattern might somehow be constructed was judged to be an unhelpful one. Having said that, although it might be right to question the significance of the Disruption Effect under these circumstances, we would still anticipate the occurrence of an Enhancement Effect.

In fact, in addition to the 'usual' factors that we might expect to have affected national curriculum test outcomes from 2016 onwards – i.e. realignment, adeptness, reallocation, and coaching – we might also anticipate an additional contestable transition impact, related to the timing of the introduction of the national curriculum. The new national curriculum was introduced in September 2014, and was tested for the first time in May 2016. However, since the national curriculum for key stage 2 spans 4 years, the first cohort to be tested would only have studied the new national curriculum (syllabus) for 2 years, having also followed the old national curriculum (syllabus) for the preceding 2 years. Only in 2018 would pupils who sat the national curriculum tests have studied the new key stage 2 curriculum in its entirety.[29] Consequently, there are threats to the interpretation of current key stage 2 test result trends lines (from 2016 onwards) arising from: inauthentic gains due to adeptness, coaching, and reallocation; and contestable authentic gains due to complex, staggered realignment effects.

It is important to emphasise that these threats are independent of the mechanisms that the Standards and Testing Agency (STA) uses to link test standards over time. That is, if these contestable transition impacts were to occur – and it seems quite likely that they would occur – then they would threaten the interpretation of trend lines *even when test standards had been linked effectively*. The sophisticated

[29] Pupils who sat key stage 2 tests in 2018 would have been taught by: year 3 teachers in their first year of teaching the new curriculum; year 4 teachers in their second year; year 5 teachers in their third year; and year 6 teachers in their fourth year, with two years worth of past papers to guide them.

techniques used by the STA to maintain test standards over time are capable of identifying and adjusting for differences in the overall demand of those tests, from one year to the next; but they are not capable of identifying and adjusting for: inauthentic gains due to adeptness, coaching, and reallocation; nor contestable authentic gains due to complex, staggered realignment effects.

Unfortunately, it is not at all clear how to respond to the conundrum that this presents.[30] Although we *would* anticipate the occurrence of contestable transition impacts under circumstances like this, the exact nature of their combined effect would remain unknown, if not unknowable. It is therefore far from obvious what guidance to provide on the interpretation of trend lines; other than a general warning not to interpret trends at face value in terms of cohort-level changes in robust understanding/valuable learning.

Having said that, in the wake of an explicit decision to apply the Comparable Outcomes principle during the early years following a transition in syllabus, assessment, and/or curriculum arrangements (e.g. for GCSEs and A levels) the ambiguity of any interpretation of result trends *would* be indisputable, and there certainly *would* be a corresponding imperative to communicate this widely.

Challenges

The biggest challenge that we face when attempting to manage the interpretation of assessment outcomes, in the wake of contestable transition impacts like the Sawtooth Effect, is our lack of detailed understanding of their operation. We have good reason to anticipate their occurrence as a consequence of syllabus, assessment, and/or curriculum transitions. However, we do not have sufficient understanding to be able to predict exactly how (or even whether) they are likely to operate in any particular transition context.

Lack of research

To some extent, our lack of understanding is due to a lack of robust research. We are beginning to make progress in this direction – a recent conference paper by Mariani (2019) provides a good example of this – but we have a long way to go. Koretz has noted a similar issue in the USA, where most of the relevant research comes from. He put it like this:

> Superintendents and commissioners generally aren't eager to have studies of possible score inflation in their systems. Trust me: asking one of them for access to their data in order to find out whether scores are inflated doesn't

[30] For instance, the idea of applying the Comparable Outcomes principle in order to mitigate the threat of Score Inflation has never been proposed in relation to national curriculum tests, in the way that it has in relation to GCSE and A level exams.

> usually get a welcoming response. So there are far fewer audits of impressive score gains on high-stakes tests than there ought to be. Nonetheless, we have enough evidence, accumulated over more than twenty-five years, to know that inflation is common and that it is often very large. (Koretz, 2017, p.53)

The North American research has established that such effects are not limited to particular assessment formats (e.g. to multiple-choice tests). It also suggests that effects might be considerably more pronounced in certain subjects than in others. For instance, Koretz has proposed that reading tests provide fewer opportunities for inappropriate test preparation "than tests in math or other content-rich subjects, such as history or the sciences" (Koretz, 2017, p.63). This helps to explain why a number of studies from North America have shown Score Inflation to be more common and more extreme in maths than reading. Having said that, other North American studies have shown that reading scores can still become severely inflated.

Beyond our general awareness of the likelihood of contestable transition impacts, we have little detailed understanding of how, and to what extent, they operate. It seems quite possible, for instance, that they may operate in subtly different ways in the UK, when compared with the USA, owing to different assessment and accountability frameworks; even when the underlying incentive structures may operate very similarly.

What seems to be particularly unclear, in relation to the UK, is the relative impact of different causal factors, especially realignment, adeptness, reallocation, and coaching. We also have only limited understanding of the timescale over which these factors might operate, a variable that is likely to interact with their relative weighting (e.g. relatively more impact from realignment early on, versus relatively more impact from coaching later on). Despite the idea of 'teaching to the test' being well-known in the UK, and often assumed to be widely practised[31], there is surprisingly little systematic research into its nature or prevalence – especially at the level of particular subject areas taught to particular year groups – with the exception of occasional observations from organisations like Ofsted (e.g. Ofsted, 2008).[32]

---

[31] "58. Many teachers reported 'teaching to the test', narrowing of the curriculum and increased pressure and workload as a result of statutory assessment and accountability." (HCEC, 2017, p.16)

[32] "113. In discussion with inspectors, although most secondary teachers recognised the importance of pedagogic skills in mathematics, they often commented on the pressures of external assessments on them and their pupils. Feeling constrained by these pressures and by time, many concentrated on approaches they believed prepared pupils for tests and exams, in effect, 'teaching to the test'. This practice is widespread and is a significant barrier to improvement." (Ofsted, 2008)

Lack of clarity

In addition to a lack of detailed understanding of the operation of contestable transition impacts, attempts to manage the interpretation of assessment outcomes across transition periods can be frustrated by a lack of clarity over the terms that we use to describe the issues at stake. Clarity is important because of the complexity of these issues, and the need to draw subtle distinctions with potentially far-reaching consequences. For instance, to be able to debate maintenance of standards cogently, it is essential to distinguish clearly between the following four reasons why cohort-level performance/attainment might rise from one year to the next:

1. (contestable) **inauthentic** performance gain – owing to a less **demanding** assessment;
2. (contestable) **inauthentic** performance gain – owing to **adeptness**, **reallocation**, and/or **coaching**;
3. (contestable) **authentic** but **unimportant** (performance and) attainment gain – owing to **realignment**; and
4. (non-contestable) **authentic** and **important** (performance and) attainment gain – owing to genuine **improvement** in teaching and/or learning.

The present paper has drawn a sharper distinction between 2 and 3 than drawn by Koretz (who has promoted the terms reallocation, coaching, and alignment); although, they may still blend into each other in certain circumstances. Yet, the distinction between an inauthentic performance gain and an authentic-but-unimportant attainment gain would seem to be fundamental; and might, for instance, invoke different intuitions on appropriate management approaches. Hence, the decision to distinguish between them as sharply as possible.

These distinctions are especially important for understanding the Comparable Outcomes principle, and how it differs from the principle that underpins traditional grade awarding practices during normal times, i.e. during periods of stability rather than transition. Previously, I have characterised the traditional logic of grade awarding as follows: if the cohort hasn't changed much, then don't expect the pass-rate to change much either (Newton, 2011). I called this the **Similar Cohort Adage**. It clearly resonates with the Comparable Outcomes principle; but it differs fundamentally in the context of its application, leading to entirely different recommendations concerning maintenance of standards.

Before Ofqual began to take a strong stance on Grade Inflation, circa 2010, we would always have assumed that any cohort-level increase in level of attainment, during a period of stability, ought to be reflected in a higher distribution of grades.[33] However, at the same time, awarding bodies have always been very pragmatic in

[33] It is a myth that A level and O level results used to be norm-referenced (Newton, 2011). It has *always* been assumed that changes in attainment ought to be reflected in changes in grade distributions.

assuming that demographically similar cohorts are quite likely to attain at similar levels; hence the Similar Cohort Adage. In effect, this renders 'no change in attainment for demographically similar cohorts' the default, null hypothesis; *unless an awarding body determines that there is a sufficiently convincing body of evidence to the contrary*. If it does deem there to be sufficient evidence, then it will sanction awarding a different profile of grades.

The Comparable Outcomes principle is quite different. It comes into play primarily during periods of transition, rather than stability.[34] Instead of treating 'no change in attainment for demographically similar cohorts' as a pragmatic null hypothesis, which is open to revision on the basis of empirical evidence, it states 'no change in grade distribution for demographically similar cohorts' simply as a matter of principle. As such, the Comparable Outcomes principle should apply even if there were to be a significant difference in level of attainment (albeit in some admittedly 'fuzzy' sense) across those demographically similar cohorts.[35]

There would seem to be at least three main reasons why the Comparable Outcomes principle and the Similar Cohort Adage are sometimes confused. The first reason is that we, the exam industry in the UK, came to embrace the Comparable Outcomes principle (during the noughties) at a time when we were losing confidence in the ability of subject experts to identify suitable grade boundary marks on the basis of expert judgement alone (see Cresswell, 2000; Baird & Dhillon, 2005; Stringer, 2012). The period from the late 2000s into the early 2010s was *both* a period of substantial qualification reform (which called for Comparable Outcomes to be applied), *as well as* a period of changing zeitgeist. This began to call for more weight to be given to the Similar Cohort Adage during grade awarding, and correspondingly less weight to be given to examiner judgement; a change that was to become more extreme during the 2010s. It is easy to see how the principle and the adage might have become confused under these circumstances.

The second reason is that prior attainment **prediction matrices** are used to apply both the principle and the adage. Even more unhelpfully, prediction matrices are sometimes referred to as the 'Comparable Outcomes technique' or the 'Comparable Outcomes method'. This is misleading because, as just observed, prediction matrices are used routinely when the Comparable Outcomes principle is *not* being applied, i.e. during periods of stability, when they are used as one of a number of tools for applying the Similar Cohort Adage. Indeed, during periods of stability,

---

[34] Although, as noted earlier, the spectre of Grade Inflation means that this is no longer quite so black and white.

[35] Consequently, the role of examiner judgement in maintaining standards according to the Similar Cohort Adage, versus the Comparable Outcomes principle, ought to be qualitatively and quantitatively different.

prediction matrices also play an extremely important role in helping to link exam standards across exchangeable qualifications offered by different awarding bodies.

The third reason is the proposal that taking a strong stance on Grade Inflation means applying the Comparable Outcomes principle indefinitely. If this were to be legislated, then the Comparable Outcomes principle would simply displace the Similar Cohort Adage; being applied both in times of transition and stability.

It is simply confusing to use the term Comparable Outcomes in relation to both a principle and a method. Conversely, the present report has referred exclusively to the Comparable Outcomes 'principle' rather than 'technique' or 'method' or even 'approach' and it is recommended that this convention should be followed more widely.[36] Both the Comparable Outcomes principle and the Similar Cohort Adage (i.e. the traditional logic of grade awarding) should be referred to quite independently of the prediction matrices technique.

[36] Or, alternatively, the 'Comparable Outcomes perspective', as originally formulated (Cresswell, 2003).

5 Summary

The Sawtooth Effect affects large-scale educational assessments, which operate under high stakes conditions, during periods of **transition**; that is, when syllabus, assessment, and/or curriculum arrangements are reformed. It is associated with the first post-reform cohort demonstrating lower quality performances, in their assessments, than the last pre-reform cohort; and with quality of performances then gradually rising over time, as teachers become ever more familiar with the new arrangements.

Over the past few years, the Sawtooth Effect has increasingly featured in discussions concerning the maintenance of standards over time within GCSEs, A levels, and other regulated assessments and qualifications in the UK. An important implication of the effect is that we should *expect* the quality of candidates' performances to drop – from the last year of a pre-reformed qualification to the first year of its reformed counterpart – owing to their teachers' initial lack of familiarity with the new arrangements. Consequently, we should also *expect* the quality of candidates' performances to rise gradually, over the next few years, as these teachers become increasingly familiar with them.

A critical question, during periods of transition, is how to manage the Sawtooth Effect: either during the grade awarding process, via grade boundary adjustments (that is, by managing assessment outcomes); or after the award of results, via a communications strategy (that is, by managing interpretations of assessment outcomes).

The Sawtooth Effect – as the term has come to be used in the UK – is a slightly 'fuzzy' concept, which embraces a number of separable effects, and which hints at a variety of causes. The present paper has introduced the more general concept of **transition impacts** – impacts upon candidates (their performances and attainments) arising from syllabus, assessment, and curriculum transitions – to describe the broader conceptual landscape that tends to be brought into view when discussing the Sawtooth Effect.

The nature of transition impacts

At the heart of the Sawtooth Effect is the idea that a change in quality of performance on the assessment, from one cohort to the next during a period of transition, might reflect *nothing more* than a change in quality of performance on the assessment; that is, it might not be attributable to a genuine change in level of attainment. One way of explaining this is that the effect of the transition is to 're-set' the first post-reform cohort to a comparatively lower baseline than the last pre-reform cohort; with the first post-reform cohort having lost the advantages that familiarity with the old arrangements had conferred upon pre-reform cohorts. For instance, the

first post-reform cohort might perform sub-optimally simply because of a lack of adeptness at recognising, navigating and responding to the demands of the new, unfamiliar assessment arrangements, which prevents them from demonstrating a quality of performance that is commensurate with their actual level of attainment. As successive post-reform cohorts benefit from the experiences of previous ones, they will become more and more adept, soon becoming able to perform optimally on the assessments.

This is just one way of explaining the Sawtooth Effect, alongside a variety of potentially complementary explanations. Most of the research and analysis into Sawtooth Effects comes from the USA. In this context, Daniel Koretz (2008; 2017) has written extensively about three alternative explanations – coaching, alignment, and reallocation. The present paper adopts and adapts these three terms, to sharpen their meaning, and to tailor them to the context of regulated qualifications and assessments in the UK. It also adds a fourth term, adeptness:

- **realignment**, i.e. as new content elements become familiar over time, teachers become better at teaching them, resources improve, and (as a consequence) learners come to learn them better;
- **adeptness**, i.e. as new assessment structures/assessment formats become familiar over time, teachers enable learners to become increasingly adept at recognising, navigating and responding to these new assessment demands, and thus more effective at demonstrating their actual levels of attainment;
- **coaching**, i.e. as new content elements/assessment formats become familiar over time, teachers begin to identify (and instruct their students in) hacks and strategies for scoring marks (without any corresponding increment in attainment);
- **reallocation**, i.e. as new content elements/assessment formats become familiar over time, teachers become better at question spotting, reallocating their instructional resources towards those elements/formats that are most frequently sampled (and away from those elements/formats that are least frequently sampled).

Reflecting upon these causal factors – in the context of maintaining assessment standards over time – it becomes clear that it is essential to distinguish between a variety of reasons why quality of performance might rise from one year to the next, even for demographically identical cohorts:

1. **inauthentic** performance gain – owing to a less demanding assessment;
2. **inauthentic** performance gain – owing to adeptness, reallocation, and/or coaching;
3. **authentic** but **unimportant** (performance and) attainment gain – owing to realignment; and
4. **authentic** and **important** (performance and) attainment gain – owing to genuine improvements in teaching and/or learning.

In this analysis, an inauthentic performance gain is one that affects *only* performance, with no commensurate change in attainment. Ignoring the possibility of Sawtooth Effects during periods of transition, the *raison d'etre* of grade awarding is:

to accommodate the possibility of a more or less demanding assessment, by setting lower or higher grade boundaries (reason 1); and therefore to enable authentic and important attainment gains or losses to be recognised (reason 4). During periods of transition, however, this process is complicated by the possibility of inauthentic performance gains due to a variety of transition-related factors (reason 2). It is also complicated by the possibility of authentic gains in attainment; albeit gains that are limited to new content elements, and that arise from a comparatively low baseline (reason 3). Reasons 2 and 3 are the stuff of Sawtooth Effects, which render the task of maintaining standards during periods of transition very complex.

Further analysis along these lines reveals that we employ the Sawtooth Effect concept when we expect the transition in question to involve: an **anticipatable** change in the quality of performance of adjacent cohorts; the significance of which is somehow **contestable**, given the presumed nature and cause of that change. The idea of contestability refers to the significance of the presumed change in quality of performance; in particular, whether it is legitimate to interpret it at face value, as indicative of a change in level of attainment. Only reason 4, above, reflects a non-contestable change, i.e. an increase in quality of performance that reflects a genuinely important rise in level of attainment. All of the other reasons reflect contestable changes. The most counter-intuitive situation occurs within reason 3, where we have an authentic change in attainment, which we choose to contest on the basis that it is an unimportant one. We invoke the Sawtooth Effect to justify its lack of importance; on the assumption that the gain in question is an inevitable, if slightly drawn-out, process of adjusting to a new syllabus, from a comparatively low baseline.

Relating this kind of analysis to wider discussions that occur during periods of syllabus, assessment, and/or curriculum transition suggests that the Sawtooth Effect is a special case within a broader category of **contestable transition impacts**. In other words, there are all sorts of causes of contestable performance changes, not all of which will necessarily trace out the characteristic sawtooth pattern of sudden drop followed by gradual rise.

Especially in the case of GCSE and A level exams, where results are used to make selection decisions, the significance of cohort-level attainment changes that are attributable to certain kinds of curriculum reform are very contestable. For instance, if the key stage 4 curriculum were to be reformed – such that fewer GCSE subjects were to be studied over the same period of time – then we would anticipate a corresponding (authentic) attainment rise in each subject area, pre-reform to post-reform. Yet, this rise would say more about the curriculum followed by those candidates, and less about their aptitudes for their chosen subjects. This would make the interpretation of any rise quite contestable. Importantly, to this example, instead of a sudden post-reform drop in quality of performance, there would be a

sudden post-reform rise. Moreover, we would not expect the impact of curriculum transition to diminish (or increase) over time, following the change.

Bearing this example in mind, it is useful to recognise that the Sawtooth Effect actually comprises two distinct effects: a **Disruption Effect**, characterised by the sudden drop; and an **Enhancement Effect**, characterised by the gradual gain. Not all contestable transition impacts will occur in exactly this manner. The Enhancement Effect can be further subdivided into: a **Recovery Effect**, which operates via realignment and adeptness; and an **Augmentation Effect**, which operates via coaching and reallocation.

Finally, it is important to emphasise that the Sawtooth Effect is ultimately a **thinking tool** rather than necessarily being manifested in either performances or in results. It is anticipatable, but not necessarily observable. Thus, by invoking the Sawtooth Effect, we typically *infer* a change in the quality of performance from one cohort to the next, rather than necessarily *observing* it directly. We presume that it is likely to be operating, even when there may be limited (or no) direct empirical evidence.

The management of transition impacts

There are two ways to manage transition impacts, including the Sawtooth Effect: either by managing assessment outcomes, during the grade awarding process, to counteract such effects; or by managing the interpretation of assessment outcomes, via a communications strategy, to counteract misinterpretations.

For GCSE and A level exams, when contestable performance/attainment changes are anticipated from one cohort to the next across a transition period, it is considered appropriate to manage assessment outcomes. This involves applying the **Comparable Outcomes** principle during the grade awarding process. Applying the Comparable Outcomes principle during the first year following a reform is tantamount to saying: even though the level of attainment of the post-reform cohort might (in some ‘fuzzy’ sense) actually have fallen, relative to the level of attainment of the pre-reform cohort, we will manage the grade awarding process as though it had remained the same (assuming that incidental demographic changes between cohorts have been controlled for). We say this, having anticipated a drop in performance/attainment – a Disruption Effect – whilst contesting its significance.

The threat of **unfairness** is considered serious enough to warrant preventative action when grading GCSE and A level exams because of how exam results are used. In both cases, results are used as a basis for selection to higher level courses, and/or to employment, and it is not uncommon for candidates from adjacent cohorts to be in competition for the same opportunities. Conversely, results from national curriculum tests are not used in this way, and the corresponding risk of unfair selection decisions does not arise. This is why the Comparable Outcomes principle

is not adopted when linking standards on national curriculum tests across periods of transition. Instead, a communications strategy management approach is adopted.

The Comparable Outcomes principle was introduced largely to manage the Disruption Effect – the sudden drop in performance/attainment following a reform – in order to be fair to candidates in the first cohort post-reform, the 'inaugural' cohort. Yet, since we would generally also anticipate an Enhancement Effect, this suggests that preventative action might also need to be taken to counteract this, too; and for exactly the same reason, i.e. fairness to candidates in successive cohorts.

Experience of, and feedback from, each new session will provide teachers (and candidates) with important insights for future sessions. These insights will naturally lead to realignment/adeptness effects. We would expect the unfamiliarity of new content elements and new assessment structures/formats to be very salient to all concerned, and for teachers to want to overcome them as rapidly as possible. Consequently, we might expect realignment and adeptness effects to be especially significant during the early years of a reformed qualification, tailing off after a certain period of time.

With experience and familiarity also comes the threat of more malign effects, too, i.e. those attributable to coaching (hacks and strategies) and reallocation (question spotting). The potential for coaching and reallocation would presumably take longer to become evident; as the predictability of new arrangements became increasingly apparent over time. Consequently, it seems likely that such effects might *step up* during the early years of a reformed qualification, rather than tail off. If this characterisation is correct, and impacts from coaching and reallocation were to extend far further into the future, then this might potentially recommend applying the Comparable Outcomes principle indefinitely. This would, in effect, be ruling out the possibility of *ever* recognising any net, non-contestable, authentic, cohort-level attainment change over time, via exam results.

It is important to recognise that there is an alternative justification for applying the Comparable Outcomes principle; that is, to counter **Grade Inflation**, rather than to counter unfairness. Indeed, to apply the Comparable Outcomes principle indefinitely might be to take a strong and defensible stance in response to the threat of Grade Inflation; even if this were at the expense of ever being able to recognise attainment trends over time via exam results. It is fair to say that there has been some lack of clarity over whether, or perhaps when, the main justification for applying the Comparable Outcomes principle is deemed to relate to fairness, or to Grade Inflation, or to both.

It is also important to appreciate that applying the Comparable Outcomes principle is a fairly crude, precautionary tactic, that we adopt in order to address a likely threat of unfairness (and/or Grade Inflation) while never being in a position to determine the 'true' state of affairs in relation to cohort-level changes in attainment over time. In

short, we apply the Comparable Outcomes principle from an unavoidable position of nescience. In the absence of any dependable knowledge concerning the occurrence of a net, non-contestable, authentic, cohort-level attainment gain or loss, we simply manage outcomes as though there had been neither (assuming that incidental demographic changes between cohorts have been controlled for).

Applying the Comparable Outcomes principle in an attempt to manage contestable transition impacts is defensible to the extent that they represent **universal** effects; in the sense of being likely to affect all, or the large majority of, candidates from a particular cohort in a similar way. This is because we manage those effects, during the grade awarding process, by adjusting grade boundary marks; and the same grade boundary marks will be applied for all candidates in a cohort. We have reasonable grounds for anticipating that adeptness and realignment impacts will be fairly universal. Reallocation and coaching seem less likely to be universal, though.

The biggest challenge that we face when attempting to manage the interpretation of assessment outcomes, in the wake of contestable transition impacts like the Sawtooth Effect, is our lack of detailed understanding of their operation. We have good reason to anticipate their occurrence as a consequence of syllabus, assessment, and/or curriculum transitions. However, we do not have sufficient understanding to be able to predict exactly how (or even whether) they are likely to operate in any particular transition context. What seems to be particularly unclear, in relation to the UK, is the relative impact of different causal factors, especially realignment, adeptness, reallocation, and coaching. We also have only limited understanding of the timescale over which these factors might operate, a variable that is likely to interact with their relative weighting (e.g. relatively more impact from realignment early on, versus relatively more impact from coaching later on).

Previously, I have characterised the traditional grade awarding principle as follows: if the cohort hasn't changed much, then don't expect the pass-rate to change much either (Newton, 2011). I called this the **Similar Cohort Adage**. It clearly resonates with the Comparable Outcomes principle; but it differs fundamentally in the context of its application, leading to entirely different recommendations concerning maintenance of standards. Conceptually, it is important to draw a distinction between the two. Applying the Comparable Outcomes principle (during periods of transition) is quite different from applying the Similar Cohort Adage (during periods of stability). Whereas the latter treats 'no change in attainment for demographically similar cohorts' as a pragmatic null hypothesis, which is open to revision on the basis of empirical evidence, the former states 'no change in grade distribution for demographically similar cohorts' simply as a matter of principle.

Finally, the present report recommends that both the Comparable Outcomes principle and the Similar Cohort Adage should be referred to quite independently of the **prediction matrices** technique. This technique is used when applying both the

principle and the adage; which means that reference to the ‘Comparable Outcomes technique’ or the ‘Comparable Outcomes method’ is misleading, and should be avoided.

6 Postscript

We began this report by considering trends in results, for maths and English, during the early years of the new GCSE qualification, from 1988 to 1995. Figure 1 gave the impression of a fairly rapid rise in results over a period of 5 or 6 years, which seemed to begin to tail off after that. A similar pattern emerged from Ben Cuff's investigation into the impact of more recent reforms to GCSEs, although his work suggested a more rapid tail-off, of just a few years (Cuff, 2016).[37]

If the steep rise in Figure 1 is consistent with the presumption that an Enhancement Effect will begin to operate following the introduction of a new qualification,[38] and if this effect can be presumed to peter out after a number of years, then this suggests an obvious method for managing outcomes: continue applying the Comparable Outcomes principle, during the early years of a reformed qualification, until the Enhancement Effect has petered out; then revert to the traditional logic of grade awarding. From this point on, hopefully, we would be able to interpret result trends over time in terms of non-contestable, authentic attainment gains or losses; at least, until the next round of qualification reform.

Unfortunately, we have already encountered a potential problem with this strategy. The Enhancement Effect can be deconstructed into two discrete effects: a Recovery Effect, which might well peter out after a few years; and an Augmentation Effect, which is likely only to get started once a few years have elapsed. In other words, it is not clear that the Enhancement Effect *would* straightforwardly peter out after a number of years. This brings us to Figure 4, below, which is an elongation of Figure 1, extending the maths and English trend lines from 1988-1995 to 1988-2016.

Before any further comment on Figure 4, it is important to stress that graphs of exam result trends over time are almost impossible to interpret definitively, post hoc; and they often give rise to appealing, yet entirely spurious, rationalisations. The biggest threat to legitimate interpretation of result trend lines is that the demographic composition of each GCSE subject cohort will change over time; and these changes will be sufficient to explain away a sizeable proportion of changes in results.

Even for GCSE maths and English, which are taken by most students each year, there can be: major changes in the demography of the cohort from one year to the next (e.g. following a large migration of candidates to international GCSE English); and there can be significant incremental changes to the demography of the cohort over longer periods of time (e.g. related to the Flynn Effect, and cohort-level IQ gains

[37] The evidence of a comparable tail-off for AS/A level was far more limited.

[38] In this example, bear in mind that delays in publishing GCSE syllabuses precluded the timely development of text books and teaching materials (Kingdon & Stobart, 1988).

over time). When the demography of the cohort changes – via changing entry patterns, changing resit patterns, changes in the balance of school types, changes in the gender balance, immigration impacts, and so on – we simply cannot interpret result trend lines straightforwardly in terms of changes in educational effectiveness.

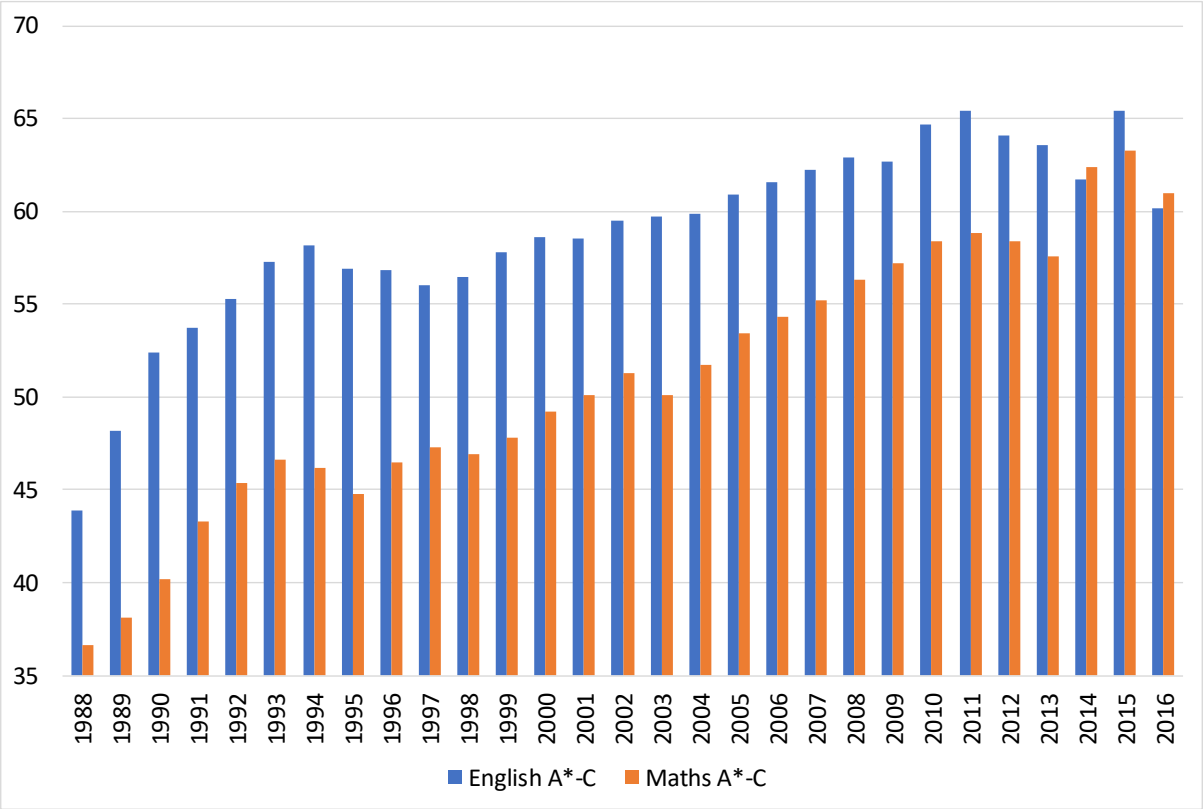

Figure 4. Cumulative percentage of GCSE subject results at grade C and above.[39]

For exactly the same reason, we cannot *necessarily* interpret rising results in GCSE English and maths, between 1988 and 1995, in terms of the Sawtooth Effect. Indeed, the general rise in results across most GCSE subjects over this period was the topic of much debate at the time; and there was speculation that results were somehow being manipulated (Newton, 1997). Roger Murphy (1996) proposed a different kind of explanation, albeit one that was equally independent of any putative Sawtooth Effect, i.e. that the observed trends could be explained by a factor as straightforward as changes in the proportion of the GCSE cohort born to families classified as being of higher social class. He presented graphs that seemed to suggest as much.

[39] This included grade A* from 1994. These data were originally prepared by the Joint Council for the GCSE/Joint Council for Qualifications, although this subset was collated by Smithers (2017). The data relate to England and Wales only up to 1995, thereafter including Northern Ireland.

Now, the point of introducing Figure 4 at this late stage is not to offer a novel explanation of the trend lines, but simply to observe that we (in England) have come to accept that at least some – if not a great deal – of the rise in GCSE results over time can be attributed to Grade Inflation, of one sort or another. The Enhancement Effect represents one sort of Grade Inflation, but there are others too.

Table 2 identifies a variety of potential causes of rising results over time. These are all causes that would operate independently of any change in the demographic composition of the cohort for each GCSE subject.[40] Of these 11 potential causes, only the 10th and the 11th are ones that might unambiguously be associated with a genuine improvement in educational effectiveness.[41] Indeed, even for these two categories, some of their underlying mechanisms would only tenuously (if at all) be associated with an improvement of this sort, i.e. more unpaid hours (teachers), more homework (learners), more home-tuition (learners).

Potential causes of rising results	Examples of mechanisms potentially underlying these rises	Transition-link
1. Systematic grading leniency	climate of expectation of rising attainment, benefit of the doubt	
2. Post-transition adeptness	teachers better at supporting learners to deal with new assessment demands	✓
3. Post-transition realignment	teachers better at teaching new content elements	✓
4. Post-transition reallocation	teachers narrow teaching of subject domain to reflect sampling approach of new assessment	✓
5. Post-transition coaching	teachers pass on strategies and hacks for circumventing the new assessment	✓
6. Changes in educational practices	redistribution of curriculum time from displaced subject areas (e.g. music)	
7. Changes in assessment practices	increased syllabus/assessment transparency, more use of reviews/appeals, more cheating	
8. Evolution of teaching/learning tools	whiteboards, internet, electronic libraries, text books, YouTube, documentaries, social media	

[40] Note that Table 2 only includes potential causes of gradual, incremental changes; consistent with the idea of Grade Inflation. It does not include potential causes of discrete, one-off changes, such as a major reorganisation of curriculum time.

[41] In theory, a case could also be made for including cause 6 in this category; but only if the displaced subjects had previously accounted for a disproportionately large amount of curriculum time.

9. Evolution of knowledge in society	academic progress, wash-back from academic progress into society more generally
10. Teachers improving their teaching	more effort, more unpaid hours, better teacher education, higher calibre entrants
11. Learners improving their learning	more effort, more homework, more home-tuition, superior meta-learning skills

Table 2. Potential causes of rising results in GCSE subjects

Relating Table 2 to Figure 4, there would seem to be more potential causes of Grade Inflation that are *not* linked to transitions (in syllabus, assessment, and curriculum arrangements) than are transition linked. In other words, however large Sawtooth Effects might turn out to be, they can only be part of the explanation of why we saw GCSE results rise relentlessly until the 2010s, when action was taken in an attempt to curb Grade Inflation. From Figure 4, the rising trend lines do seem to be a little steeper from 1988 to 1993/4; which might hint at a stronger Sawtooth Effect during these years. Again, though, we need to be extremely cautious not to over-interpret such trend lines, as there have been all sorts of changes along the way.[42]

This Postscript began by considering the case for attempting to manage the Sawtooth Effect during the early years following a transition, by applying the Comparable Outcomes principle; after which we might then revert to the traditional logic of grade awarding. It ends by concluding that any such proposal needs to be set alongside the observation that Grade Inflation is caused by a variety of factors; many of which are not transition linked. Consequently, in addition to the challenge of not knowing exactly how long any Sawtooth Effect would need to be managed, we simultaneously have to confront the challenge of whether, and if so then how, to manage non-transition-linked causes of Grade Inflation.

[42] For instance, an important change occurred with the introduction of the Mandatory Code of Practice for the GCSE, which came into full operation in 1994. It coordinated grade awarding practices across awarding bodies, with requirements such as the following from paragraph 121: “The Chairman of Examiners must recommend grade boundaries to the Chief Executive, providing evidence in support of boundary recommendations that lead to proportions of candidates within grades which differ significantly from those of previous years. The Chief Executive must decide whether or not to endorse the Chairman of Examiners’ recommendations.” (SCAA, 1994, p.26)

7 References

Baird, J., & Dhillon, D. (2005). *Qualitative Expert Judgements on Exam Standards: Valid, but inexact*. Internal report RPA 05 JB RP 077. Guildford: Assessment and Qualifications Alliance.

Baird, J., Caro, D., Elliott, V., El Masri, Y., Ingram, J., Isaacs, T., Pinot de Moira, A., Randhawa, A., Stobart, G., Meadows, M., Morin, C., and Taylor, R. (2019). *Exam Reform: Impact of Linear and Modular Exams at GCSE*. OUCEA/19/2. Ofqual/19/6506/1. Coventry: Office of Qualifications and Exams Regulation.

Baird, J., Hopfenbeck, T.N., Ahmed, A., Elwood, J., Paget, C. & Usher, N. (2014b). *Predictability in the Irish Leaving Certificate Examination. Working paper 1: Review of the literature*. Oxford: Oxford University Centre for Educational Assessment.

Baird, J., Hopfenbeck, T.N., Elwood, J., Caro, D. & Ahmed, A. (2014a). *Predictability in the Irish Leaving Certificate (OUCEA/14/1)*. Oxford: Oxford University Centre for Educational Assessment.

Cannell, J.J. (1988). Nationally normed elementary achievement testing in America's Public Schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practice*, 7 (2), 5-9.

Cresswell, M.J. (2000). The role of public exams in defining and monitoring standards. In H. Goldstein & A. Heath (Eds.). *Educational Standards* (pp. 69-104). Oxford: Oxford University Press for The British Academy.

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

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.

Cuff, B.M.P., Meadows, M., & Black, B. (2019). An investigation into the Sawtooth Effect in secondary school assessments in England. *Assessment in Education: Principles, Policy & Practice*, 26 (3), 321-339.

Department for Education (2016). *National curriculum assessments at key stage 2 in England, 2016 (provisional)*. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/549432/SFR39_2016_text.pdf

Holmes, S., Khan, A., Zanini, N. & Black, B. (2020). *Predicting Predictability: Investigating question paper predictability and the factors that influence this through a question prediction exercise*. Coventry: Office of Qualifications and Examinations Regulation.

House of Commons Education Committee (2017). *Primary Assessment. Eleventh Report of Session 2016–17* (HC 682). Available from: https://publications.parliament.uk/pa/cm201617/cmselect/cmeduc/682/682.pdf

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed). *Educational Measurement* (4th edition) (pp.17-64). Washington, DC: American Council on Education/Praeger.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50 (1), 1-73.

Kingdon, M. & Stobart, G. (1988). *GCSE Examined*. Sussex: The Falmer Press.

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.

Koretz, D.M., Linn, R.L., Dunbar, S.B. & Shepard, L.A. (1991). *The Effects of High-Stakes Testing On Achievement: Preliminary findings about generalization across tests*. California, LA: University of California Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Linn, R.L. (2000). Assessments and accountability. *Educational Researcher*, 29 (2), 4-16.

Linn, R.L., Graue, M.E., & Sanders, N.M. (1990). Comparing state and district test results to national norms: The validity of the claims that "Everyone is above average." *Educational Measurement: Issues and Practice*, 9 (3), 5-14.

Mariani, E. (2019). *Measuring and Correcting the 'Sawtooth Effect' in a First Award*. Paper presented at the 20th Annual Conference of the Association for Educational Assessment – Europe. Lisbon, Portugal. 13th-16th November.

Murphy, R.J.L. (1996). Like a Bridge over Troubled Water: realising the potential of educational research. *British Educational Research Journal*, 22 (1), 3-15.

Newton, P.E. (1997). Examining standards over time. *Research Papers in Education*, 12 (3), 227-248.

Newton, P.E. (2010). Thinking about linking. *Measurement: Interdisciplinary Research and Perspectives*, 8 (1), 38-56.

Newton, P.E. (2011). A level pass rates and the enduring myth of norm-referencing. *Research Matters*, Special Issue, 2, 20-26.

Nisbet, I. and Shaw, S.D. (2019). Fair assessment viewed through the lenses of measurement theory. A*ssessment in Education: Principles, Policy & Practice*, 26 (5), 612-629.

Ofsted (2008). *Mathematics: Understanding the Score: Messages from inspection evidence*. London: Office for Standards in Education.

Pollitt, A. (1998). *Maintaining Standards in Changing Times*. Presented at the 24th Annual Conference of the International Association for Educational Assessment. Barbados, West Indies. May, 1998.

School Curriculum and Assessment Authority (1994). *Mandatory Code of Practice for the GCSE*. London: School Curriculum and Assessment Authority.

Shepard, L. (1990). Inflated test score gains: is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*, 9 (3), 15-22.

Shepard, L. (1997). *Measuring Achievement: What does it mean to test for robust understanding?* Princeton, NJ: Policy Information Center, Educational Testing Service.

Smithers, A. (2017). *GCSE Trends 1988-2016 & Prospects for 2017*. University of Buckingham, Buckinghamshire: Centre for Education and Employment Research.

Stringer, N.S. (2012). Setting and maintaining GCSE and GCE grading standards: the case for contextualised cohort-referencing. *Research Papers in Education*, 27 (5), 535-554.

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/20/6710/1