Cambridge Assessment

Research *Matters*

Issue 31 / Spring 2021

Cambridge Assessment

Proud to be part of the University of Cambridge

Established over 160 years ago, Cambridge Assessment operates and manages the University's three exam boards and carries out leading-edge academic and operational research on assessment in education. We are a not-for-profit organisation.

Citation

Articles in this publication should be cited using the following example for article 1: Shaw, S. D., & Nisbet, I. (2021). Attitudes to fair assessment in the light of COVID-19. *Research Matters: A Cambridge Assessment publication, 31,* 6–21.

Credits

Reviewers: Filio Constantinou, Carmen Vidal Rodeiro, Sylvia Vitello, Joanna Williamson (all Research Division, Cambridge Assessment) and Andrew Watts (Wolfson College, University of Cambridge)

Editorial and production management: Anouk Peigne, Research Division, Cambridge Assessment

Additional proofreading: Alison French

Cover image: Getty Images

Design: Lente Artemieff, Louise Stubbings

Cambridge Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team:
Research Division, researchprogrammes@cambridgeassessment.org.uk

If you need this document in a different format contact us, telling us your name, email address and requirements and we will respond within 15 working days.

All details are correct at the time of publication in April 2021

Research *Matters* / 31

A Cambridge Assessment publication

Contents

Research *Matters* / 31

A Cambridge Assessment publication

Foreword

And then came pandemic. The Research team at Cambridge Assessment turned from the established research programme not only because of disruption to work, but because the organisation needed to focus on how best to ensure continued delivery of the important public goods which assessment and qualifications supply to individuals, society and the economy. We needed to supply grades and certification in both academic and vocational qualifications. And being researchers, we marshalled all the existing evidence which could inform the process of giving grades in the absence of exams and made that available for decision-making. We placed researchers in operational teams to support awarding. We contributed to national policy work and convened special teams to interpret processes as they were applied for the first time. We used prior research to help build the guidance for teachers and the data portal for them to submit data. We ran processes which created data to drive the awarding. And despite the lack of the normal stream of information from examinations, and despite lack of access to schools, we knew that we had to gather information to interpret the impact of the pandemic—to understand the quality and fairness of the alternative arrangements which had been put in place, the patterns of "learning loss" and the impact of both disruption and the policy response. And while the domestic scene required constant attention, the organisation also was providing certification to over 170 countries, all with their own patterns of disruption and distress. The articles in this edition of *Research Matters* are the first articles from our effort to understand what occurred, and to understand what we need to put in place to introduce greater resilience into assessment and qualifications. We now go into a second year of alternative approaches in England, and a mixture of exams and alternatives internationally. We go into the summer conscious of the burst of grade inflation in 2020, and with a recognition that the cohorts of summer 2021 have suffered greater learning loss than those in 2020. We will continue to trace the impact of the pandemic—particularly the progression of learners and the vexed issue of standards. That will take time. But simultaneously—and quickly—we are extracting from our review of events the key characteristics of future arrangements which will meet the complex demands we place on assessment—which include but are not limited to: recognition of learning, supporting admissions and progression, specifying the key content and outcomes of learning programmes, and signalling the quality of arrangements for learning.

Tim Oates, CBE Group Director, Assessment Research and Development

Research *Matters* / 31

A Cambridge Assessment publication

Editorial

Welcome to the new online version of *Research Matters*. The first issue in our new format is a special issue devoted to research relating to the COVID-19 pandemic. In the UK the exams were cancelled in summer 2020, and in all four UK nations students were eventually awarded the better of a Centre Assessed Grade (CAG) and a grade produced by an "algorithm" which statistically standardised the CAGs with the dual aims of compensating for differences among schools in how harsh or generous their estimated grades were, and minimising overall grade inflation. The experience raised, and continues to raise, fundamental questions about fairness, standards, reliability, the meaning of grades, the purpose of assessment—and others. In this special issue we touch on many of these issues.

In our first article, Stuart Shaw and Isabel Nisbet discuss the fairness of methods of awarding grades in the absence of exams, in the light of the approach set out in their timely and widely praised new book on fairness of assessment, written before the pandemic (see details on page 100).

In the second article, Tom Benton shows that the grade inflation in England resulting from the decision to abandon "the algorithm" is explicable in terms of giving the benefit of the doubt in circumstances where there is more doubt than usual.

As is often the case, more media attention was given to GCSEs and A Levels than to Vocational and Technical Qualifications (VTQs), but solutions for grading VTQ students were needed too—and a single approach was not possible because of the variety of assessment models. In our third article, Sarah Mattey gives a brief guide to what happened with VTQs in summer 2020.

Our fourth article, by Melissa Mouthaan and colleagues, compares and contrasts policies on curriculum, pedagogy and assessment in the four home nations in the first six months of the pandemic. This is followed by Gill Elliott's overview of how the pandemic will have affected learners in England in all the different school age groups.

The word "unprecedented" has understandably been used many times since the start of the pandemic, but of course exams have been disrupted before. In our final article, Gillian Cooke and Gill Elliott draw on documents from the Cambridge Assessment Archives to see if the response to past disruptions can provide some perspective for our current situation.

I hope you find the contents interesting, and our new format easier to access and navigate. Do send any feedback to researchdivision@cambridgeassessment.org.uk

Tom Bramley Director, Research Division

Attitudes to fair assessment in the light of COVID-19

Stuart Shaw Cambridge Assessment International Education and **Isabel Nisbet** Faculty of Education, University of Cambridge

"Exams are the fairest way to assess what students know and can do." (Office of Qualifications and Examinations Regulation, 2020f)

"We were looking at … fairness across a whole population … but acknowledging from the outset that it would not be anything like as accurate as exams." (Roger Taylor, Chair, Ofqual; Parliament (UK), 2020a)

Introduction

Was the approach proposed for calculating exam grades in summer 2020 fair? Were the grades eventually awarded (after policy changes) fair? What is a fair arrangement for 2021? These questions have been at the heart of debate in the United Kingdom (UK) in the light of COVID-19. The language of fairness has been uppermost for all involved: assessment professionals, teachers, students, parents, journalists and politicians. After schools were closed in the spring of 2020 and the decision was made not to proceed with summer exams, it was judged unfair to deny students the grades they needed to progress to the next stage in their lives. The task was to find a fair way to award grades in the absence of exams.

The approaches developed in all four parts of the UK—and the Republic of Ireland—were thought by the regulators and ministers to be the fairest possible, but in the event the grades initially awarded were widely decried as "unfair" and instead all the UK countries switched to awarding Centre Assessed Grades (CAGs).[1] The result was not only significant grade inflation (compared to previous years) but unequal treatment of different subjects and groups of candidates which the regulator for Wales described as showing "some unfairness" that should not be repeated in future (Qualifications Wales, 2020 (a)).

In this article we shall briefly recall the conceptual map of "fairness" that we have offered elsewhere (Nisbet & Shaw, 2020) and outline received views of assessment fairness before 2020. We shall then discuss five challenges to those received views raised by the COVID-19 experience, particularly in the UK.

Conceptual mapping: fair assessment

Previously, we have distinguished senses of fair that can be confused in discussions about educational assessment (Nisbet & Shaw, 2019; 2020). Four main senses are:

1 This was qualified (in slightly different ways in different countries) by allowing the student the "best of" the Centre Assessed Grade or the grade that they had already received through the statistical approach used originally and then dropped (e.g., Ofqual, 2020e).

- *Formal:* denoting accuracy or the appropriate application of a rule or design.
- *Implied contractual:* something is fair if it meets the legitimate expectations of those affected.
- *Relational—treating (relevantly) like cases alike*: discrimination is fair if it is based on relevant considerations and unfair if it is based on something else, such as the candidate's race or gender (relational fairness is key to much discussion of assessment fairness).
- *Retributive:* an outcome is fair if it is an appropriate reward (or penalty) for what has gone before. In this sense, saying that a candidate's grade was fair would mean that the candidate was thought to deserve it.

Each of these senses of fair can be contrasted with unfair, and there is no doubt that our emotional reaction to unfairness is often keenly felt, as every parent of a young child will affirm. This article is concerned with attitudes to assessment and it is necessary to consider throughout whether the attitudes described are primarily negative reactions to unfairness or approbation of fair practices.

It is often assumed that fairness applies only to candidates. But there are others whose interests may also be at stake, for example, candidates' peers (who did not take the test), users of the assessment outcomes (e.g., employers or universities) or society at large.

Received views of fair assessment

A broad consensus on fairness has developed among assessment professionals and academics. The received view, enshrined in authoritative documents such as the North American *Standards for Educational and Psychological Testing* (hereafter North American *Standards*), sees fairness as an absence of unfairness, with unfairness shown by construct-irrelevant variance in assessment outcomes. Unfairness so understood can be identified, in arrears, by "differential functional analysis" and prevented, in advance, by "universal design", avoiding bias. The consensus view focuses almost entirely on relational fairness and arguably does not do justice to the retributive senses of fair or the importance of "legitimate expectations".

The language of fairness in assessment is often (and, in our view, mistakenly) confined to groups rather than addressing fairness to individuals (e.g., Isaacs et al., 2013; cf. Nisbet & Shaw, 2020, pp.20–23). However, the 2014 issue of the North American *Standards* does extend the concept to individuals. It portrays fairness as a fundamental right of all individuals and subgroups in the test population: a fair test "reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population" (AERA et al., 2014, p.50).

There is no question that discussions of fairness of the grades awarded in 2020 applied that concept to individuals as well as groups. Indeed, the Chair of the Office of Qualifications and Examinations Regulation (Ofqual) suggested to Members of Parliament (MPs) that grades calculated using statistics were perceived as unfair because "the level of accuracy that was fundamentally possible … was too low to be acceptable to individuals" (Parliament (UK), 2020a, Q998).

Challenges and questions from the COVID-19 experience

The impact of the pandemic on education (including assessment at all levels) and on the lives of students, families and educators, has raised the profile of some aspects of assessment fairness and questioned some of the assumptions about fair assessment inherited from pre-COVID-19 times. In this article we identify and discuss five such challenges to thinking about fair assessment, which are:

1. Fairness and public attitudes—the role of "felt" fairness
2. Fair assessment in context—"opportunity to learn"
3. Equality versus desert
4. Fairness and maintaining standards over time
5. Relationship of fairness to validity, reliability and comparability

1. Fairness and public attitudes—the role of "felt" fairness

An independent review commissioned by the Scottish Government (Priestley et al., 2020) reported on the attitudes of young people to the changing situation regarding the summer exams in 2020. The review report depicts the announcement of the cancellation of exams—without clarity about what evidence was to be used in their place—as provoking a "visceral reaction" (Priestley et al. 2020, p.37), with "students crying and screaming" (Priestley et al. 2020, p.37). The report criticised "a lack of appreciation, by key bodies throughout the process, that the issue of perceived fairness to individuals might become a toxic political issue if not handled with sensitivity and forethought" (Priestley et al., 2020, p.42).

Talk of "perceived" or "felt" fairness in educational contexts is not uncommon nowadays (e.g., Nisbet & Shaw, 2020, pp.35–36). However, cries of perceived unfairness were louder and more frequent than usual in summer 2020. Exams may have had their faults, but they were known and planned-for. Suddenly, what was expected was removed and replaced by uncertainty and rumour. This felt unfair, in the legitimate expectations sense. But was it really unfair? To what extent is felt fairness the same as actual fairness? Can a perception of (un)fairness be wrong? And how have perceptions of felt fairness shifted during the COVID-19 experience?

Taras (2002) has advanced the notion that "students perhaps have the right to demand coherent and logical educational processes that are not detrimental to their learning" (p.501). We know from relevant research that students embrace complex, and sometimes contradictory notions of fairness, being more inclined to identify instances of unfairness (Sambell et al., 1997; Orr, 2010). Flint and Johnson (2011) have identified criteria for fair assessment from the perspective of the student which cover several of the senses of fair identified here, the strongest influence being legitimate expectations, violated when the assessment is a nasty shock.

After the (to some) unsettling news of the cancellation of the exams in 2020, the next "nasty shock" in 2020, across the countries of the UK, was the award of grades calculated by the use of a statistical model. In many cases these grades were lower than the estimates by candidates' schools, and this led to accusations of unfairness from students, their parents and teachers.

The infamous[2] "algorithm" (Stewart, 2020) used in England to calculate grades combined rank orders from teachers with information on the historical performances of schools and the prior attainment of candidates. Broadly similar models were developed in the other UK countries. The model preferred for England was one of a range of possibilities which were thoroughly analysed and the subject of consultation, including with school leaders (Ofqual, 2020a). However, calculated grades which seemed fair in aggregate later seemed unfair to teachers and school leaders when they saw the implications for their own schools and their own students.

Was it justified to perceive the use of statistical models as unfair? The algorithm was an attempt to use statistics to achieve as much relational fairness as possible, and the regulators carried out technical analyses to look for potential unfairness as understood in the received view, namely construct-irrelevant differences between some groups or categories of student and others (for England see, for example, Ofqual (2020d)). One potential relational unfairness that was identified was that the statistical approach could not be validly applied to centres with very small numbers of entries in a given subject, and so in those cases the (more generous) CAGs were to be used. This appeared to benefit unfairly the students who took these subjects and their schools, which were often independent schools. In England particularly,[3] this meant that a technical analysis of fairness of assessment raised wider questions of social justice and the role of private education in the class system.

After the change of policy and the award of CAGs, the regulators for England and Wales both claimed that the grades calculated using the algorithm were less unfair in this regard (favouring schools with small subject cohorts) than were the CAGs eventually used (Parliament (UK), 2020a, Q946; Qualifications Wales, 2020 (a)). However, that was not how it felt—there was much more discussion of the differential effect of the algorithm on the calculated grades in different types of school than there was about the (perhaps greater) differences in the grades that were eventually used.

We suggest that another aspect of the calculated grades that felt unfair was that they reflected calculations of probability, which felt unjustified when applied to individuals. As the Chair of Ofqual explained to MPs, "[I]f you have 1,000 students that have, for example, an 80% chance of getting an A grade, they would regard themselves quite reasonably as A-grade students. What we were doing in effect was recognising that, in a normal year, 200 of those students would fail to get their A grade" (Parliament (UK), 2020a, Q945). In the absence of the exam, the 200 students out of 1,000 who were denied their A grade by the algorithm felt aggrieved. Arguably, that feeling was reasonable, given the absence of evidence about their own work that fed into the decision. As the Chair of Ofqual remarked, "this whole process ultimately feels unfair to the individual, because they have not had the appropriate degree of agency" (Parliament (UK), 2020a, Q981).

2 https://www.theguardian.com/politics/2020/aug/26/boris-johnson-blames-mutant-algorithm-for-exams-fiasco

3 This form of relational unfairness was less contentious in the other UK countries, where there are proportionately fewer independent schools.

2. Fair assessment in context—"opportunity to learn"

National examinations, such as GCSEs, AS and A Levels in parts of the UK, can become rituals of national life. Each year, attention may be focused on some aspect of the exams themselves—their content or their difficulty, or whether girls did better than boys—but there has been much less discussion, at exam time, of what went before the test was taken—the teaching and learning experienced by different groups of students. Differences in the educational experience of students are often discussed but seldom linked with perceptions of the fairness of exams.

This state of affairs contrasts with the USA, where popular and specialist discussion of the fairness of tests frequently refers to differences in students' opportunity to learn the content being assessed. In the 2014 edition of the North American *Standards*, Standard 3.19, confined to "settings where the same authority is responsible for both curriculum and high-stakes decisions based on curriculum mastery", includes: "[E]xaminees should not suffer permanent negative consequences if evidence indicates that they have not had the opportunity to learn the test content" (AERA et al. 2014, p.72).

If the test's requirements go beyond the curriculum it is intended to test then, the authors of the North American *Standards* would argue, that is unfair. It is unfair in the retributive sense, as the test content does not match what it is supposed to cover and candidates who have studied that curriculum but done badly in the test will not deserve their low mark. And if some groups of candidates have had the opportunity to learn about the content domain of the test while other groups have not, that is also unfair in the relational sense.

In the USA, this line of thinking was influenced by the leading legal case of *Debra P. v. Turlington*, which was considered no fewer than four times by the courts between 1979 and 1984. It concerned a "functional literacy test", introduced by the State of Florida as a requirement for a High School Diploma. Black students performed very badly in the test compared to their white counterparts, and it was argued that students who failed the test might not have been taught the test content in earlier years when schools were racially segregated. Successive courts introduced a concept first labelled "curriculum validity", which denoted the fit between what students had been taught and the content of the test. The label was later changed to "instructional validity", as the contrast was not with what students were supposed to be taught but what they were actually taught. And the Circuit Court was clear—"If the test covers material not taught to the students, it is unfair".[4]

In the UK, the experience of COVID-19 in 2020 has brought to the fore concerns about the loss of teaching and learning time by students. During the period in spring/summer 2020, when schools were closed for most students, there was concern about differences in the quality and quantity of remote teaching and learning available to students and in their ability to make use of it, which was affected by family circumstances and access to technology. And from autumn 2020, there has been considerable local variation in school attendance because of COVID-19, as well as evidence from Scotland of a "strong relationship between pupil disadvantage and school attendance levels" (EPI/Nuffield, October 2020; see also Sibieta, 2020).

4 644 F. 2d 397 (5th Cir. 1981), p.4.

When considering the implications for fair assessment, we need to distinguish between, on the one hand, concerns that all or most students may have missed out on some of the learning that would normally be expected for the assessment; and, on the other hand, concerns that there are wide differences between the amounts of learning lost by different groups of students. Both raise issues of fairness linked to opportunity to learn. But, as Nick Gibb MP stated to MPs, if all students had lost broadly the same amount of teaching and learning time, it would be possible to compensate for this—at least to some extent—by reducing the mandatory content of papers and adjusting grade boundaries. However, there would remain a problem, described by the Minister as "differential unevenness and unfairness" (Parliament (UK), 2020c, Q1122), if there were wide variation between the amount and quality of teaching experienced by different groups, to an extent not acknowledged before.

Of course, there have always been differences between students' opportunities to learn, which are not the students' fault and hence unfair in a retributive sense. These could reflect different qualities of teaching, poverty, family circumstances, ill-heath, or access to additional help from parents or tutors. And more fundamentally, the differences in students' talents and abilities—described by Nagel as "the injustice of the smart and the dumb" (Nagel, 1979, p.104)—is arguably itself unfair in a retributive sense. However, the visible differences in learning lost by school students in 2020 as a result of COVID-19 have struck home with policymakers in a way that "normal" differences have not.

As we shall discuss later, different UK countries have responded in different ways to the problem of COVID-19-induced differences in opportunity to learn. The Welsh Government concluded that these differences made it unfair—in a relational sense—to hold exams at all in summer 2021 (Welsh Government, 2020). In England, Government and the regulator argued in December 2020 that fairness—perhaps, we would comment, in the legitimate expectations sense—required exams to be held, but that COVID-19-related differential learning loss might be reflected in some kind of record "alongside the [exam] grade" (Parliament (UK), 2020d, Q58). Despite those differences of response, however, there did seem to be a shared concern about differential opportunities to learn, as a result of COVID-19, and about the implications for (relational) fairness. It is too early to tell whether this will have an impact on attitudes to differential opportunity to learn after COVID-19.

3. Equality versus desert

In *Is Assessment Fair?* (Nisbet & Shaw, 2020) we discussed the philosophical roots of concepts of fairness. We depicted a balance between notions of (some kind of) equality—which were reflected in the relational senses of fairness discussed above—and those of desert—linked to the retributive sense (Nisbet & Shaw, 2020, p.108).

As we have seen, the received view of assessment fairness focuses almost entirely on the relational sense. In contrast, desert requires that each candidate gets the grade he or she deserves. This can be viewed using two different perspectives. Focusing on the test itself, a fair assessment accurately measures the relevant knowledge or skill of the candidate, and discussion of fairness in this sense often uses the language of "accuracy" and "reliability". A wider perspective sees the assessment in context, with a fair assessment outcome seen as "deserved" because of the hard work that the student has done or because it matches some other evidence of the student's ability. How has the balance between equality and

desert been reflected in attitudes to national qualifications in the shadow of COVID-19? We suggest that there have been three phases in the public discussion in the UK during 2020, and that the balance has been differently struck for each. During that time, there have also been marked changes in our knowledge about COVID-19 and our experience of living with it.

The first phase was the preparation for the awarding of grades in summer 2020, following decisions to cancel examinations. In this phase, relational fairness, based on equality, was paramount. Regulators and examination authorities across the UK were required by governments to develop an approach which would maintain standards and apply those standards in the same way across the country. Intensive work was done to develop and then evaluate possible statistical models for calculating grades to achieve those kinds of equality. Where any model risked advantaging or disadvantaging particular school types—for example, independent schools with small subject cohorts—that was seen to be a reason for concern.

As the Chair of Ofqual acknowledged, while a model for calculating grades could provide fairness "across a whole population … it would not be anything like as accurate as exams" (Parliament (UK), 2020a, Q945). Without evidence about the work of individuals—whether from an exam or some other source—it was not possible to design an approach that would match each individual's grade to evidence of their own work. In our terms, desert was bound to take second place to equality.

The second phase was the reaction to the calculated grades in summer 2020. The immediate hostile reaction was based on desert at the individual level. Although at a national level the calculated grades were more generous than the previous year (Ofqual, 2020b), individuals who were disappointed by their grades felt that they deserved better. Where the grades estimated by their teachers had been "downgraded" by a statistical method, students were aggrieved that they had been disadvantaged by the application of an algorithm without reference to evidence about them as individuals. The intention was that individual claims of unfairness could be pursued through the appeals system, but the momentum of discontent made that unsustainable. Discontent about desert led to the change of tack in each of the countries of the UK to award CAGs.[1]

At the time that decision was made, it was already known that CAGs would be significantly more generous than the grades awarded in 2019 and earlier years. It was subsequently claimed by the regulators that CAGs displayed more relational unfairness between centre types than did the calculated grades (Qualifications Wales, 2020(a); Ofqual, 2020h). However, after the immediate furore had died down, information released showing relational unfairness in the award of CAGs attracted little comment, other than some concern that students who entered university with over-generous grades might find it difficult to cope. Why this change in attitudes to fairness? Part of the explanation may lie in "outrage fatigue"—when the row about calculated grades had led to a change, many more students were able to secure their university places, and the press had moved on to the next news story. However, we suggest that there were two lines of thought about fairness which can be traced in attitudes at this stage of the public debate. The first was that complaints about desert dominated over concerns about equality and that predominance persisted, even when the relational unfairness of the CAGs was evident. It is possible that students, teachers and parents saw the grades awarded simply as a (deserved) reward in the particular circumstances of 2020, rather than a token whose

worth could be compared with other years.[5] A second underlying belief may have been that (strict) comparability between centres and across subjects was not as important as supporting students after the hardship of the lockdown period and enabling them to progress to the next stage of their education.

The third phase was the debate in autumn and winter 2020/21 about whether to hold summer examinations in 2021. This started with considering whether it was desirable or practicable to plan to hold traditional-style examinations. Initially, the UK countries answered this question differently, with England the last country to retain a commitment to hold summer examinations, but in January 2021, in the context of renewed closures of schools for most pupils, English ministers reluctantly cancelled plans for exams and commissioned work to develop alternative assessments (Department for Education, 2021).

Rarely have exams been so praised as in their absence. Giving oral evidence to the UK Parliamentary Select Committee on Education, Gavin Williamson MP emphasised that "[t]he best and fairest form of assessment is a proper form of examination" (Parliament (UK), 2020b). In saying that, he was echoing the view of the regulator cited at the beginning of this article. But what did the Minister mean by "fair"?

We suggest that he was using considerations of both equality and desert. In comparison with the documented unevenness of alternatives to exams—whether calculated grades or CAGs—he may well have looked favourably at the tried and tested methodologies for security and standardisation of grades for national exams and regarded those as better guarantors of relational fairness (based on equality) across the cohort of students. He may also have felt that an exam provides some record of the individual student's work which can be referred to when judging whether the grade awarded to a student was deserved. The lack of such an individualised record was one of the main reasons for criticising the calculated grades (based on an algorithm) as unfair. When questioned by MPs, the Minister of State for School Standards (in England) said: "Having exams is the fairest way to enable students to demonstrate, *through their own work* [emphasis added], what they know and what they have achieved" (Parliament (UK), 2020d, Q88).

In our discussion of opportunity to learn, we have described the perceived significance for fair assessment of the considerable variations within individual countries of the UK, in the amount and quality of learning time lost by students because of COVID-19. In contrast to the view of ministers in England, Welsh ministers concluded that differential loss of learning meant that fairness—based, we would comment, on equality—required the cancellation of exams in 2021 (Welsh Government, 2020). In England, at the time that ministers remained committed to holding exams, there were suggestions that assessment standards could be varied across regions, to reflect differences in lost learning. The reply by ministers and the regulator was that such a practice would be (relationally) unfair, because it could not take into account differences within the regions concerned. In the words of the Minister of State, "those sorts of adjustments would bring their own unfairnesses" (Parliament (UK), 2020d, Q80). Fairness as equality has come back to the fore in this third phase of debate.

5 We owe this observation to Joanna Williamson, Cambridge Assessment Research Division.

4. Fairness and maintaining standards over time

In 2007, the Education Secretary, Ed Balls, announced that he was establishing an independent regulator of qualifications in order to end the "old and sterile debate" about exam standards being "dumbed down" (BBC News, 2007). And in the subsequent legislation to establish Ofqual, its duties included ensuring that regulated qualifications "indicate … a consistent level of attainment (including over time) between comparable regulated qualifications".[6]

What is the link between fairness and maintaining standards over time? It seems unarguable that in some circumstances it would be unfair—in the relational sense—if two people competed for one university place on the basis of grades they had achieved in different years, and the standards required for achieving these grades were different. It is less clear why it would be unfair if standards changed over a longer period—say, 10 years—although there might be an argument that relational fairness required students in these different years to have parity of esteem for the quality of their work. There might also be (relational) unfairness to later students if their grades were less valued by prospective employers than those of students who obtained their grades in times past, although as students would presumably normally be compared with their contemporaries, rather than with earlier generations, that does not seem a very strong argument. Another argument might be that if more students got higher grades, the mark scheme might not recognise very high achievers in a way that they deserved, and that would be unfair.

Whatever the justification, emphasis on maintaining standards over time had not diminished by March 2020, when the decision was made to cancel summer exams because of COVID-19. The Department for Education's (DfE) Direction to Ofqual included: "Ofqual should ensure, as far as is possible, that … the distribution of grades follows a similar profile to that in previous years" (DfE, 2020a). Governments in other parts of the UK gave similar instructions.

The approach to calculated grades developed by the exam authorities and regulators across the UK was developed with a view to keeping outcomes "broadly in line" with those of previous years, but also seeking to minimise (construct-irrelevant) differences between outcomes across centres. Ofqual said that although the calculated grades would be "slightly higher" than in previous years, the "currency of the qualifications for progression" would not be "undermined" (Ofqual, 2020b).

However, the grades awarded after the changes of policy in summer 2020—largely based on assessment by centres—were not "broadly in line" with standards in previous years, but markedly more generous. This was evident in all the UK countries, but a striking instance was the award of high grades at A Level in Wales, where "at cumulative grade A*-A revised results in 2020 were 43.7%, compared to 27% in 2019" (Qualifications Wales, 2020(b), paragraph 6.1). In England, where students were given the opportunity to take exams in November 2020 that were not available in the summer, Ofqual stated that for reasons of fairness they would "work … with exam boards to carry forward the generosity from summer 2020 grades" in the November exams (Ofqual, 2020g). Ofqual subsequently decided, again citing reasons of (relational) fairness, to carry forward the generosity of the 2020 grades to summer 2021. The regulator argued that, while the standards required for

6 Apprenticeships, Skills, Children and Learning Act 2009, S128 (2)(b)(i).

particular grades would not be the same as those before 2020, "students in 2021 [would] have as much chance of getting a grade A or a grade 4 as they did in 2020" (Stacey, 2020).

Reflecting on these developments, it is clear that the initial predominance of concern about standards over time was overtaken in August 2020 by the negative responses to calculated grades, and the anger that some students had not achieved the grades they and their teachers thought they deserved. Although the underlying moral arguments were seldom articulated, we suggest that attitudes implied a judgement that the moral case for supporting students who had had a particularly tough time—perhaps a form of high-level fairness as desert—was thought more important than maintaining standards over time. Tellingly, the Irish Education Minister, after observing the turbulent debate in the UK, included in a public statement about the Irish Leaving Certificate: "We have … lessened the importance placed on the historic national standards" (Department of Education, Ireland, 2020b).

In our opinion, the principle of maintaining standards over time was probably always more closely linked to confidence and to what Scottish ministers called "the integrity and credibility of the qualifications system" (Priestley et al., 2020, p.6) than to arguments about fairness. As we have seen, justification of the principle in terms of fairness is possible, but requires a context to make clear why differences matter—for example, in competition for the same course or job. Once that context is past—for example, when students have been accepted for their university courses or have obtained a job—cries of unfair about different standards are less persuasive. For example, students whose grades were subject to harsher standards in 2019 than their post-COVID-19 successors were not strong voices in discussions about the fairness of the grades awarded in 2020. In our view, that is understandable—should they feel aggrieved that those who (unlike them) had their teaching and learning disrupted by a pandemic were assessed using more generous standards?

No doubt the "dumbing down" argument in the early 2000s detracted from confidence in qualifications, and it was plausible to say that that could be countered by visible maintenance of standards over time. Such thinking survived into the early months of 2020 but, by the end of the summer, public attitudes tolerated outcomes which clearly breached the principle of maintaining standards over time. If an attempt is to be made in the future to peg back grade standards to pre-COVID-19 levels, in order to restore confidence, there will almost certainly be cries of unfair by the first students who are subjected to the harsher standards than their immediate predecessors.

5. Relationship of fairness to validity, reliability and comparability

According to the received view of many assessment theorists, validity refers to the interpretation of the results of an assessment, with valid interpretations being significant, useful and appropriate (AERA et al., 2014, p.11). Fairness is seen as a fundamental aspect of validity, depicting the "validity of test score interpretations for intended use(s) for individuals from all relevant subgroups" (AERA et al., 2014, p.219). We have previously argued that fairness (in most but not all of the senses identified in this article) is a necessary but not sufficient condition for validity, although this will depend to a large extent on how validity is conceptualised (Nisbet & Shaw, 2020). In any event, validity and fairness are closely linked, and both are central to public confidence in tests and their

outcomes, even if the language of fairness is more familiar in public discourse than is that of validity.

Fairness is closely tied to the concept of reliability and any threat to reliability will call the fairness of the test into question. And the received view of assessment fairness, focusing on relational fairness and the absence of construct-irrelevant bias, also invites a close link between fairness and comparability. For example, a chemistry test which does not have comparable outcomes for individuals or groups of candidates with relevantly similar knowledge of chemistry will be thought unfair.

In the absence of exams, what kind of assessment can lead to interpretations which are valid and reliable? And what is the link with fairness? Reflecting on the experience in Scotland in 2020, the Priestley review sought to shift attention away from the question of "how suitable the algorithm was for the task"—arguably, a very narrow concept of validity—to "whether the task was operationalised in a valid way" (Priestley et al., 2020, p.43). It concluded that the interpretations drawn from results were invalid—and unfair—because they were not based on evidence of "the effort and achievement" of individual students (Priestley et al., 2020, p.43).

The validity of exam grades depends on their interpretation. In judging the validity of the substitutes for exams in 2020, it helps to distinguish three possible interpretations of the resultant grades:

(a) As a measurement of the relevant knowledge and skill demonstrated at the time of the assessment (or previously). This formulation is characteristically used for assessments used for summative purposes.

(b) As an indicator of the stage of learning reached and the appropriate learning to follow. This typically characterises assessments—often in the classroom—used for formative purposes.

(c) As an indicator of the potential of the candidate for something in the future, such as a university course or a job.

In normal times, an exam sat in the summer would be primarily understood in terms of (a) above—as a measure of attainment—with a loose link to (c)—as an indicator of potential—although aptitude tests of a different kind are sometimes used specifically for that purpose. All three purposes can be distinguished from a fourth, described by Ofqual (referring to the use of the algorithm) as aiming to "reflect the grades students would have been most likely to achieve if teaching and learning had continued and they had taken their exams as normal" (Ofqual, 2020c).

This interprets the grades as a (counter-factual) judgement of what would have been awarded, at a different date from the date of the judgement, in circumstances which did not happen. As such it is unverifiable by direct evidence (unless the circumstances happen after all) and becomes a probability judgement.

It is perhaps understandable that a statistical approach was used for this fourth purpose. Arguably, an interpretation of the grades awarded as representing probability judgements would be valid, and if the approach used was relationally fair to different categories of students and schools, it might be seen as fair. But it seems less persuasive that the use of the algorithm was valid for interpretation (a) or (c) as applied to the individual student.

One of the options which Ofqual put to ministers following the decision to cancel exams was assessment for a "teacher certificate" encouraging a different kind of interpretation (Parliament (UK), 2020a, Q948). If it was largely based on assessments by teachers, the interpretation might allow for a wider degree of variability in the circumstances and the judgements made than was expected for interpretation (a). However, that in itself would raise issues of fairness for a cohort of students who could expect to receive a traditional grade (legitimate expectation) and who would be competing for university places and jobs with other cohorts who had such grades (relational fairness).

The Welsh Government has proposed an approach for 2021 involving "teacher-managed assessments" (Welsh Government, 2020). They seem to be aiming for the grades awarded to validate interpretation (a) and to be seeking to achieve relational fairness by involving an element of externality and some form of moderation. It is too early to judge whether the interpretation of the grades generated by this approach will satisfy the requirements of validity in relational fairness.

Turning to reliability, the Priestley review suggested that varied approaches to estimating grades in Scotland detracted from their reliability. Teachers' estimates were clearly "subject to variation (in the types of evidence available, the processes followed for internal moderation and the support given by local authorities)" (Priestley et al., 2020, p.12). Smith (2003) defines the most appropriate standard for reliability for classroom purposes in terms of "sufficiency of information" and asks the question: "Do I have enough information here to make a reasonable decision about this student with regard to this domain of information?" (p.30). That question does seem relevant to assessments in the classroom used as information sources in the absence of exams.

In our view, the COVID-19 experience in the UK should prompt reconsideration of validity, reliability and fairness in relation to substitutes for traditional exams. In considering validity, it will be necessary to distinguish the four interpretations which we have identified. Depending on which is intended, the importance of some concepts of validity and reliability may be secondary to some of the concepts discussed in this article, notably opportunity to learn. Bonner (2013) has suggested that "measurement validity may be a secondary concern" (p.87) within the situated, locally embedded nature of classroom assessments. This comment may also have to apply to "teacher-managed assessments" in the absence of exams.

Conclusions

To unpack attitudes to fair assessment in the context of COVID-19, we have revisited the different senses of fairness that are relevant to educational assessment and outlined the received view of fair assessment, before the pandemic. We have outlined five challenges to that view brought by the COVID-19 experience. We now set out a few generalised conclusions.

The impact of felt (un)fairness reaches across the other challenges. The COVID-19 disruption has revealed areas of contention hitherto rarely discussed in the UK in the context of exams, notably the differences in the teaching and learning experienced by different groups of students and the notion of fairness based on opportunity to learn. This has come to the fore in discussions of the fairness of exams in the UK, with particular

reference to 2021. Whether this is a temporary or longer-term shift in attitudes remains to be seen.

We have argued that notions of fairness are derived from root concepts of equality (equal outcomes for candidates who are equal in construct-relevant respects), which resonates with the relational sense of fairness, together with that of desert, reflected in the retributive sense of fairness. We have traced the changing balance between the two in attitudes shown during three phases of the debate prompted by COVID-19, with the emphasis shifting from equality to desert and back to equality again. In our view, an account of assessment fairness must allow for considerations of both equality and desert. Models which focus exclusively on one and ignore the other may lose touch with public attitudes.

Prior to the onset of COVID-19, the principle of maintaining standards over time remained steadfast and it is still a statutory duty of Ofqual. Ultimately, however, the grades awarded in 2020 were strikingly more generous than those awarded in previous years and this was tolerated by professional, political and public attitudes. We have discussed why this was so and suggested an implicit underlying moral argument. However, it remains to be seen whether relegation of that principle in public and policy priorities will be temporary or longer-lasting.

The final challenge related to the link between fairness and traditional measurement concepts of validity, reliability and comparability. One key question prompted by COVID-19 has been the interpretation of grades awarded, which is relevant to a validity argument, and we have distinguished several possible interpretations, some of which seem less applicable if grades are awarded without reference to a traditional exam.

In the light of the public outcry in 2020, it is very unlikely that attitudes in the UK would tolerate an approach in the near future based entirely on probability estimates using a statistical model. In developing alternatives to exams, there will be a need to take account of a range of evidence from schools and colleges as evidence of achievement by individuals. The demands of validity, reliability and fairness will need to be reconsidered in that context. The debate about whether traditional psychometric concepts like validity should be re-conceptualised for the purposes of classroom settings is not a new one (see Smith and others in a Special Issue of *Educational Measurement: Issues and Practice* (2003)). However, the pandemic experience affords a new opportunity for re-invigorating the discussion.

Accusations of unfairness raised in attitudes to assessment in the light of COVID-19 need to be taken seriously by the assessment profession, regulators and governments and cannot be assumed to be a temporary phenomenon. In the words of evidence to the Priestley Commission, "each statistical point on the graph is an individual young person" and an approach to grading cannot be accepted if it "creat[es] an overall perception of fairness but fails to deliver actual fairness for individuals" (Priestley et al., 2020, p.27).

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Apprenticeships, Skills, Children and Learning Act 2009, UK Public General Acts, S128 (2009). https://www.legislation.gov.uk/ukpga/2009/22/section/128

BBC News. (2007, September 26). *Watchdog to tackle exam standards.* http://news.bbc.co.uk/1/hi/uk_politics/7013033.stm

Bonner, S. M. (2013). Validity in Classroom Assessment: Purposes, Properties, and Principles. In J. H. McMillan (Ed.), *Sage Handbook of Research on Classroom Assessment* (pp.87–106). Sage.

Debra P. v. Turlington, 644 F.2d 397 (5th Circuit 1981). https://case-law.vlex.com/vid/644-f-2d-397-597132630

Department for Education. (2020a). *Direction to the Chief Regulator of Ofqual about GCSE, AS and A Level qualifications.* https://www.gov.uk/government/publications/direction-issued-to-the-chief-regulator-of-ofqual

Department of Education. (2020b). *Minister Foley announces details of Calculated Grades model for Leaving Certificate 2020 and Minister Harris confirms the provision of 1,250 additional places in certain high-demand programmes in higher education institutions* [Press release]. https://www.education.ie/en/Press-Events/Press-Releases/2020-press-releases/PR20-09-01.html

Department for Education. (2021, January 6). *Education Secretary outlines plans to support young people* [Press Release]. www.gov.uk/government/news/education-secretary-outlines-plans-to-support-young-people

EPI/Nuffield. (2020, October). *Analysis: School attendance rates across the UK since full reopening*. Educational Policy Institute. https://epi.org.uk/publications-and-research/school-attendance-and-lost-schooling-across-england-since-full-reopening/

Flint, N. R., & Johnson, B. (2011). *Towards Fairer University Assessment. Recognising the Concerns of Students.* Routledge.

Isaacs, T., Zara, C., Herbert, G., Coombs, S. J., & Smith, C. (2013). *Key concepts in Educational Assessment* (1st ed.). Sage.

Nagel, T. (1979). *"The Policy of Preference", in Cambridge University Press (Eds), Mortal Questions. Cambridge University Press.*

Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, *26*(5), 612–629. https://doi.org/10.1080/0969594X.2019.1586643

Nisbet, I., & Shaw, S. D. (2020). *Is Assessment Fair?* Sage.

Ofqual. (2020a). *News story: Ofqual GCSE and A level consultation outcomes and autumn exam series proposals*. GOV.UK. https://www.gov.uk/government/news/ofqual-gcse-and-

a-level-consultation-outcomes-and-autumn-exam-series-proposals

Ofqual. (2020b). *Letter to Heads of centre - Summer 2020 results.* GOV.UK. https://www.gov.uk/government/publications/letter-to-heads-of-centre-summer-2020-results

Ofqual. (2020c). Fairness in Awarding. *The Ofqual Blog*. https://ofqual.blog.gov.uk/2020/08/06/fairness-in-awarding/ (retrieved 18/11/20).

Ofqual. (2020d). Research analysis: *Awarding GCSE, AS & A levels in summer 2020: interim* report. GOV.UK. https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report

Ofqual. (2020e). Changes to awarding of GCSE, AS and A level: guide for teachers, student, parents & carers: summer 2020. GOV.UK. https://www.gov.uk/government/publications/changes-to-awarding-of-gcse-as-and-a-level-guide-for-teachers-student-parents-carers-summer-2020

Ofqual. (2020f). *Ofqual welcomes DfE announcement on 2021 exams* [Press release]. https://www.gov.uk/government/news/ofqual-welcomes-dfe-announcement-on-2021-exams

Ofqual. (2020g). Setting standards in the autumn series. *The Ofqual Blog*. https://ofqual.blog.gov.uk/2020/10/19/setting-standards-in-the-autumn-series/

Ofqual. (2020h). News release: *Summer 2020 outcomes did not systemically disadvantage students.* GOV.UK. https://www.gov.uk/government/news/summer-2020-outcomes-did-not-systemically-disadvantage-students

Orr, S. (2010). Collaborating or fight for the marks? Students' experiences of group work assessment in the creative arts. *Assessment & Evaluation in Higher Education, 35*(3), 301–313.

Parliament (UK). (2020a). Transcript of oral evidence to the Select Committee on Education by Roger Taylor, Chair, Ofqual (and others). https://committees.parliament.uk/committee/203/education-committee/publications/oral-evidence/

Parliament (UK). (2020b). Transcript of oral evidence to the Select Committee on Education by the Rt Hon Gavin Williamson, Secretary of State for Education (and others). https://committees.parliament.uk/committee/203/education-committee/publications/oral-evidence/

Parliament (UK). (2020c). Transcript of oral evidence to the Select Committee on Education by the Rt Hon Nick Gibb MP, Minister of State for School Standards (and others). https://committees.parliament.uk/event/2411/formal-meeting-oral-evidence-session/

Parliament (UK). (2020d). Transcript of oral evidence to the Select Committee on Education by Dame Glenys Stacey, Interim Chief Regulator, and the Rt Hon Nick Gibb MP, Minister of State for School Standards. https://committees.parliament.uk/oralevidence/1365/pdf

Priestley, M., Shapira, M., Priestley, A., Ritchie, M., & Barnett, C. (2020). *Rapid Review of National Qualifications Experience 2020*. Final Report, September 2020. Faculty of Social Sciences, University of Stirling. https://www.gov.scot/binaries/content/documents/

govscot/publications/independent-report/2020/10/rapid-review-national-qualifications-experience-20202/documents/rapid-review-national-qualifications-experience-2020/rapid-review-national-qualifications-experience-2020/govscot%3Adocument/rapid-review-national-qualifications-experience-2020.pdf

Qualifications Wales. (2020, October 12). Statistical release: *Variation in GCSE, AS and A level qualification results in Wales, Summer 2020*. https://www.qualificationswales.org/english/publications/variation-in-gcse-as-and-a-level-qualification-results-in-wales-summer-2020/

Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, *23*(4), 349–71.

Sibieta, L. (2020, October). *School attendance rates across the UK since full reopening.* EPI and Nuffield Foundation. https://epi.org.uk/wp-content/uploads/2020/10/UK-school-reopening-attendance-EPI.pdf

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice*, *22*(4), 26–33.

Stacey, G. (2020, December 3). *Exams and assessments in the months ahead: How Ofqual will set standards for exams and assessments to be fair to students.* GOV.UK. https://www.gov.uk/government/speeches/exams-and-assessments-in-the-months-ahead

Stewart, H. (2020, August 26). Boris Johnson blames 'mutant algorithm' for exams fiasco. *The Guardian.* https://www.theguardian.com/politics/2020/aug/26/boris-johnson-blames-mutant-algorithm-for-exams-fiasco

Taras, M. (2002). Using Assessment for Learning and Learning from Assessment. *Assessment & Evaluation in Higher Education*, *27*(6), 501–510.

Welsh Government. (2020, November 10). Wales' approach for qualifications in 2021 confirmed by Education Minister Kirsty Williams [Press release]. https://gov.wales/wales-approach-qualifications-2021-confirmed-education-minister-kirsty-williams

On using generosity to combat unreliability

Tom Benton Research Division

Introduction

This article concerns how we might think about acting in a situation where we know that, for whatever reason, an assessment is less reliable than normal. At the time of writing, the instance in mind is the cancellation of all exams in England in summer 2020 and the ultimate decision that grades should (largely) be awarded based upon teacher estimates in the form of Centre Assessed Grades (CAGs) instead.[1] However, the same principles may apply in other scenarios such as where, for practical reasons, it may be necessary to shorten the length of some assessments or alter their composition in some other way.

This article will not attempt to address all of the varied aspects of fairness that may be relevant to such a situation. A review of issues of fairness, such as whether qualification grades should simply be fair in terms of treating all students equally, or whether they must provide a "fair" reward for the effort put in by students is provided by Shaw and Nisbet (2021, this issue). Similarly, we will not be considering all of the different factors that may have affected different students to a greater or lesser extent over the past year and how assessment can be made fair in the light of this. Rather, this article will focus on a single issue—the positioning of grade boundaries. We will explore whether psychometrics can provide a rational approach to setting grade boundaries in a situation where an assessment system is less reliable than would typically be the case.

Once we know that the reliability of an assessment is lower than normal, the natural human inclination is to try to ensure that students do not lose out. Reduced reliability will increase the chances of a student ending up with a better grade than they would have achieved with a more reliable assessment. However, usually the greater concern is that it will also increase the chances of students being awarded a worse grade than would otherwise have been the case. Since the reduced level of reliability is due to circumstances that are beyond students' control, it seems unfair that some of them may end up with lower grades than they would have achieved under ordinary circumstances. As such, we may wish to ensure that assessments are graded leniently so as to reduce the number of students getting a lower grade than they would have done, while accepting this will mean more getting a higher grade than they would have done. It is clear that we would not know which particular individuals benefit and which are disadvantaged by the unreliability.

A straightforward way to incorporate leniency into assessment is during the awarding process when we select grade boundaries. Setting grade boundaries sufficiently low reduces the difficulty of achieving each grade and may mitigate the impact of reduced assessment reliability on students. This article considers this issue in more detail. In particular, the aim is to discover exactly how much generosity is needed during awarding

1 A discussion of why we believe such estimates may be less reliable than full-length exams will be provided in a later section.

to combat increases in unreliability, and how this depends upon the strategy to awarding that we take. Interestingly, we will show that, in some respects, the final distribution of grades in summer 2020 was similar to what might be expected from a logical application of giving students the benefit of the doubt from a position of uncertainty about how they would have performed in real exams.

True scores, true grades, observed scores, awarded grades, reliability

Before going any further, we need to define our terms. From now on, we define a student's *true score* (or *true mark*) as the score that they would achieve on a 100 per cent reliable assessment (i.e., an assessment that was very long and had completely reliable marking). True scores could be transformed to follow any statistical distribution, so, for the purposes of this article, we define *true grades* as being defined by students' percentile rank rather than by their score. For example, we might decide that the top 25 per cent of students should be awarded grade A (or above). Note that the concepts of "true scores" and "true grades" are entirely notional to help us think about issues relating to reliability. In reality we never observe the true score of any student. However, for the purposes of this research we imagine that we could in order to think about issues such as student misclassification.

Observed scores are the scores that are actually assigned to students including the measurement error inherent in any real assessment. Note that error, in this context, does not necessarily need to mean "mistake". Rather, it refers to anything that may lead to changes in students' scores between replications of an assessment process. For example, certain items will suit some students more than others, meaning that students' performances usually vary between different test versions. Similarly, there may be differences in the professional judgement of examination markers, meaning that re-marking may lead to changes in students' scores. We define *awarded grades* or *observed grades* as the grades awarded to students based on their observed scores.

Following usual practice in classical test theory, we define *reliability* as the squared correlation between true and observed scores. This is equivalent to the percentage of the variance in observed scores attributable to variance between students' true scores. It is also mathematically equivalent to the expected correlation between observed scores on two parallel tests (very roughly speaking, these are tests made to the same specification so that, in particular, they consist of equally difficult items and are of equal length).

Ensuring that no student is disadvantaged

In a perfect world we would ensure that, even with reduced reliability, absolutely no student ends up with a lower grade than they deserve. In the midst of a crisis regarding an assessment, statements to this effect are often released by the organisations responsible for them. What is usually left unsaid is that we would, under the circumstances, perhaps accept some students being awarded higher grades than they deserve in order to ensure that no student misses out. This initial section considers exactly how generous we would need to be in order to achieve this goal.

Figure 1 provides an imaginary example to introduce these ideas further. The plot shows the relationship between simulated true marks for 1,000 students on the horizontal axis and their observed marks from a sub-optimal assessment process on the vertical axis. For the purposes of this simulation, both true marks and observed marks were simulated to be between 0 and 100, to have a mean of 50 and a standard deviation of 16 (see footnote[2]). The correlation between true and observed scores was simulated to be 0.8—that is, the reliability of observed scores was 0.64 (=0.8^2).

Next, to begin with, imagine that we have initially decided that the top 25 per cent of students should be awarded grade A. For example, we may have initially considered that, despite the increased unreliability, we would like to preserve the grade distribution from previous years so that the grades awarded to students were indistinguishable from those awarded under the normal assessment process. As such, we identify grade boundaries on both the true and observed marks that achieve this. These are represented in Figure 1 by the solid black lines and are 61 out of 100 in each case.

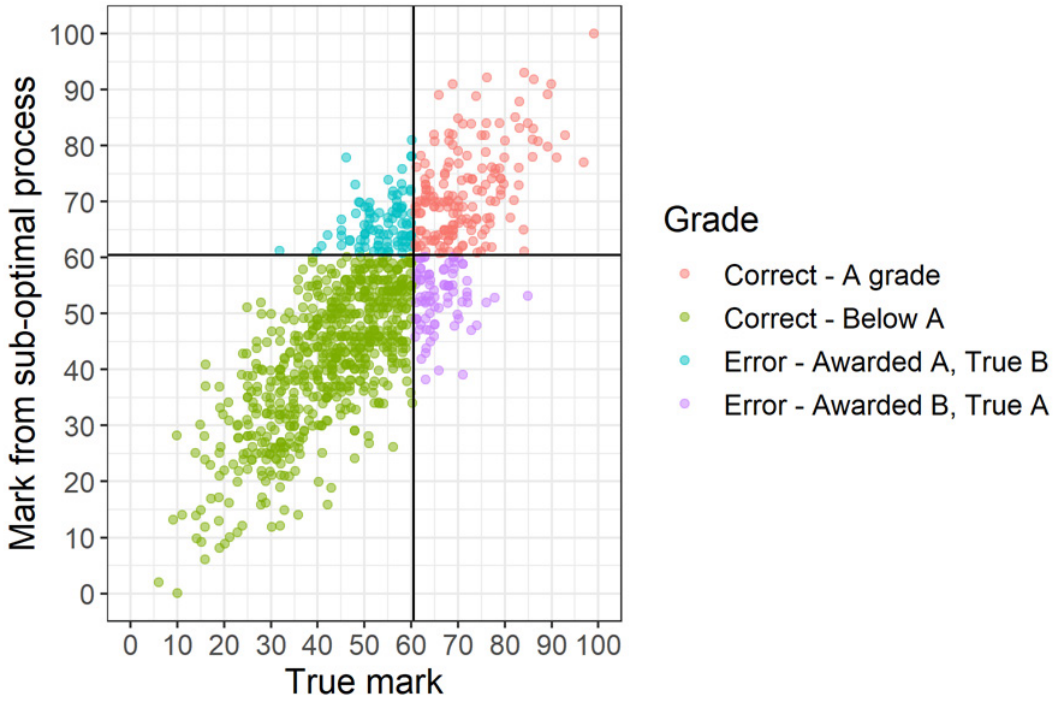


Figure 1: Relationship between true marks and grades, and marks and grades from an unreliable assessment if we choose to award the same percentage of top grades.

Suppose next that, having realised the possibility of students missing out on the grade they deserve, we decide to lower the grade boundary to eradicate this risk. This process is illustrated in Figure 2. For the purposes of this figure, the grade boundary on observed scores has been lowered to the point where no student is awarded a grade lower than they deserve. This has required lowering the grade boundary from 61 out of 100 to 38 out of 100. Furthermore, in total, only 464 students are now awarded the correct grade. The good news is that all 254 students that deserved a grade A have been awarded this (the red points in Figure 2). However, this has come at a cost. Lowering the grade boundary

2 These are reasonably typical score distributions for examinations. For example, see Table 1 of Wheadon and Stockford (2010).

has led to 536 students being awarded a grade A when, according to their true scores, they did not deserve this (the blue points).

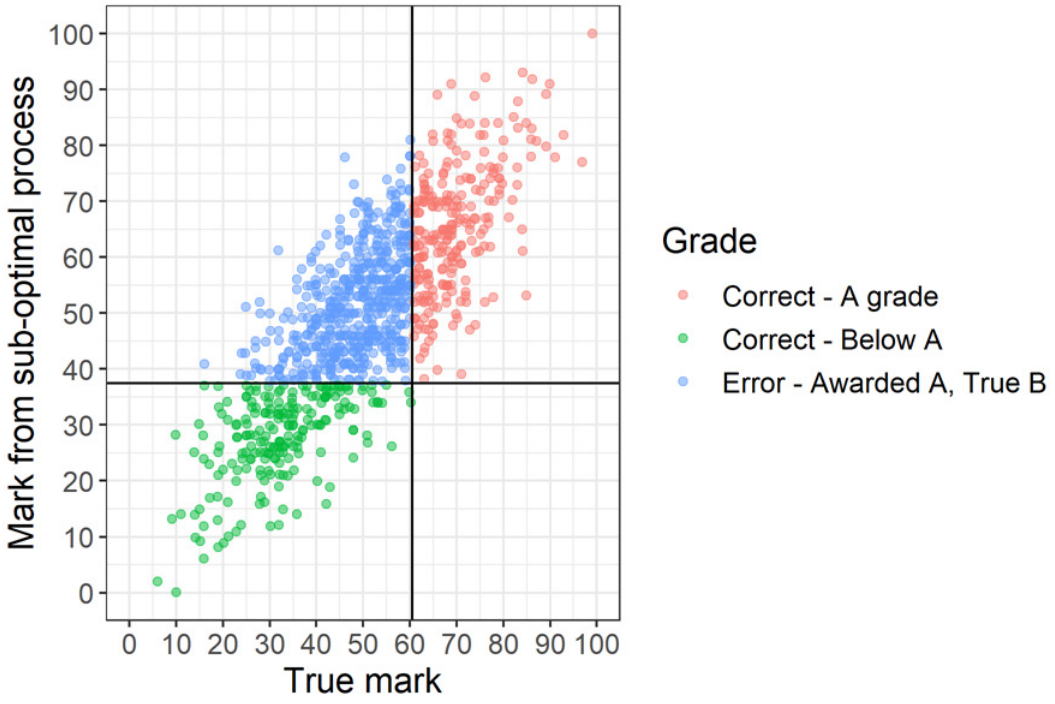


Figure 2: Relationship between true marks and grades, and marks and grades from an unreliable assessment if we lower the grade boundary for the unreliable set of marks.

Figure 2 demonstrates the difficulty of ensuring no student is disadvantaged by unreliability. Measurement error can affect students both positively and negatively and to differing amounts. In the presence of unreliability, with a sufficiently large number of students, there will always be at least one person whose observed score is far below their true score. Lowering grade boundaries far enough to mitigate this risk in its entirety would require an enormous dose of generosity. This in turn risks undermining the meaning of having achieved a particular grade, devaluing the currency of the qualification in question, and undermining public confidence. As such, ensuring that absolutely no student is disadvantaged is not something that can realistically be achieved through the awarding process alone.

Figure 2 imagines that we are trying to ensure that no student is disadvantaged relative to their true score. However, true scores are never observed and so in reality "an assessment will never be perfectly reliable" (Wheadon & Stockford, 2010, p.32). As such, "it is inevitable that some candidates with true scores in one grade will, on some occasions, achieve a score just outside that grade" (Wheadon & Stockford, 2010, p.32). With this in mind, while we might accept that we cannot lower grade boundaries to a point where no student achieves a grade below the one they deserve, we may be able to lower them to an extent where the risk of a student being under-graded by an abnormally unreliable assessment is no higher than the risk we accept under normal circumstances. This is the focus in the next part of this article.

Reliability of various assessment processes

Before we can consider this issue, we need some idea of the reliability of assessments under normal circumstances. Our focus will be on A Levels and will be on the reliability of entire qualifications rather than the individual examinations (components) which comprise these. Although several existing pieces of research have published data on the reliability of individual examination components (e.g., Bramley & Dhawan, 2010; Hayes & Pritchard, 2013; Wheadon & Stockford, 2010) very few of these have attempted to calculate reliability at whole qualification level. In order to estimate the reliability of a typical A Level, we use the following pieces of information:

- At present, a typical A Level will require students to complete at least three separate examination components which have roughly equal weight.
- The median reliability (Cronbach's alpha) of individual A Level examination components is 0.83 (Bramley & Dhawan, 2010, Table 2.5). Note that since Cronbach's alpha is often considered a lower bound for the true reliability of a component (see Hayes & Pritchard, 2013), this means that the above estimate of the reliability of a full A Level may be slightly conservative.
- Analysis of A Level components taken with OCR in June 2019 reveals the median correlation between examination components taken as part of the same A Level qualification (subject), is 0.64 (based on 173 pairs of components).[3]
- Combining the above facts with a simplified version of the Wang and Stanley formula for calculating composite reliability (He, 2009) gives us:

$$\textit{Typical A Level reliability} = \frac{3\alpha + 6\rho}{3 + 6\rho} = \frac{3 * 0.83 + 6 * 0.64}{3 + 6 * 0.64} = 0.925$$

  Where α is the expected reliability of each individual component, ρ is the expected correlation between components, "3" is the number of components, and "6" is the number of ways in which three components can be paired (respecting the direction of pairings).

As such, we will use a reliability of 0.925 as representing the reliability of a typical A Level. This estimate is a little lower than the estimate of 0.963 for a full-length Mathematics A Level provided by Benton (2014). This may reflect the reduced amount of assessment in reformed A Levels,[4] the use of the (more conservative) Cronbach's alpha in the present calculations, or the focus on a typical A Level subject (using medians) rather than exclusively on Mathematics.

To provide a richer context for the calculations in subsequent sections, we will also estimate the "reliability" of an alternative, and less reliable, assessment method—forecast scores. The word "reliability" is placed in inverted commas in the previous sentence to represent the fact that, as discussed earlier, we are defining reliability as the squared

3 Restricted to components taken by at least 200 students.

4 Since recent reform, A Levels no longer combine results from assessments taken over a period of two years and, instead, focus on (fewer) examinations taken in a single session.

correlation between any given measure and true scores from examinations. This is not quite the same as the usual definition of reliability (i.e., how much would forecast scores change if we replicated the forecasting procedure) but it is the most relevant definition in the context of this article.

The word "forecast" is used above as the estimates of reliability (below) are derived from an analysis of A Level forecast grades submitted to OCR by teachers in May 2014 ahead of students completing their exams. They may differ from the predicted grades that were sent to UCAS (Universities and Colleges Admissions Service) earlier in the academic year as part of the university application process (see Gill & Benton, 2015, for further details). Our interest in forecast grades stems from the fact they are the best information we have available to say anything about the reliability of teacher estimated grades such as CAGs. Although we have collected data on CAGs, data on exam achievement of the same students in the same subjects is not available (because exams were cancelled). As such, comparing forecast grades and actual exam grades is the best source of information we have. Further discussion of the relationships between different types of teacher estimated grades is provided in McManus et al. (2020).

With the above discussion in mind, we estimate the reliability of forecast scores as follows:

- For OCR A Levels in summer 2014 the median polychoric correlation[5] between forecast and actual A Level grades was 0.82 (across 57 A Level specifications).
- Given the calculations earlier and adjusting for the unreliability of actual A Level results, we use a standard correction formula (Spearman, 1904) to estimate that the correlation between forecast scores and true qualification scores is 0.85 (=0.82/√0.925).
- Thus, the estimated "reliability" of forecast scores is 0.72 (=0.85^2).

Note that forecast scores, as such, have never existed—only forecast grades. However, we use the concept here to help us think about the likely reliability of procedures based upon teachers estimating the likely future achievement of their students (such as was done for the production of CAGs in summer 2020).[6] That is, the estimate gives an idea of the confidence with which teachers can estimate the future performance of their students. The idea of a continuous forecast score is useful as it allows us to apply concepts relating to the positioning of grade boundaries within this context.

The idea that results based on this type of teacher assessment are less reliable than those based on formal examinations has some support in the existing research literature. For example, McManus et al. (2020) showed that forecast grades tend to have considerably lower correlations with future undergraduate and postgraduate achievement at university than actual A Level grades.

5 We use a polychoric correlation to adjust for the fact that coarse scales (such as grades) are likely to be less highly correlated than fine scales (such as marks). The median Pearson correlation between forecast and actual grades was slightly lower at 0.77. Only A Levels with at least 200 entrants were included in calculations. See Gill and Benton (2015) for more details about the accuracy of forecast grades.

6 See Ofqual (2020).

It should be noted that the reliability estimate here only concerns a particular type of teacher assessment. The estimate does not necessarily apply to other types of internal school assessment such as where teachers mark pieces of standardised coursework or controlled assessment and these marks are then moderated. Previous research (Benton, 2016) has shown that, in at least some circumstances, such approaches to assessment are likely to be just as reliable as formal examinations.

Possible choices regarding standards in the face of added unreliability

To recapitulate, we wish to explore different approaches to setting grade boundaries in the situation where we have been forced to switch from a form of assessment with a reliability of 0.925 (approximate reliability of a full A Level qualification) to a form of assessment with a reliability of 0.72 (approximate reliability of forecast scores). We will consider three different approaches to setting grade boundaries:

- **Retain the grade distribution.** That is, we decide that despite the change to the reliability of qualifications we should retain the grade distribution we have always had. For the purposes of calculations in this section we will assume the grade distribution should match the overall grade distribution for A Levels from 2019 (JCQ, 2019).
- **Ensure no increase in the percentage of students awarded a grade below their true grade.** That is, we aim that the number of students awarded a grade below their true grade is no higher for the less reliable assessment method than for the normal assessment method. We abbreviate this method as the No Added Disadvantage strategy.
- **Maximise accuracy.** That is, we choose grade boundaries to maximise the percentage of students awarded a grade equal to their true grade. This leads to slightly different results from method 1. This will be explored more as we display the results.

To ease the calculations in this section, we will assume that all scores (both true and observed) follow normal distributions. Given that any set of scores could be transformed to follow a normal distribution (i.e., converted to normalised scores) this is not such a huge assumption. Since, after transformation, scores could be expressed on any scale, grade boundaries will be expressed in terms of the proportion of students at each given grade or above. Note that, based on these assumptions, all of the calculations in this section were completed entirely by applying mathematical formulae (i.e., there was no need to simulate data). Examples of the R code used to complete calculations are provided in the appendix.

To begin with, we focus upon setting an A* boundary. Results for other grades will be presented later but this provides a good starting point to illustrate the concepts in this research. Note that we define an A* as referring to the top 7.8 per cent of students.

Figure 3 shows the estimated misclassification rate at A* for a full-length A Level depending upon the percentage of students we award the grade to. The green curve shows how the proportion of students awarded a grade that is too high (i.e., an A* when their true grade is below A*) changes with the proportion awarded A* overall. If very few people

are awarded A* (the left of the chart) then hardly any will be over-graded. As the number awarded A* increases so does the proportion of students that are awarded a grade that is too high. The blue line is of more interest. This shows the percentage of students under-graded as the grade boundary moves. Harsh grading (to the left of the chart) results in a greater percentage under-graded than lenient grading (to the right of the chart). The red curve denotes the total level of misclassification—the sum of under-grading and over-grading.

The vertical purple line represents where we would position grade boundaries for the full-length qualification. That is, we award A* to exactly 7.8% of students. Because this assessment is not perfectly reliable (no assessment is) some students are misclassified. Specifically, given an assessment reliability of 0.925, 1.6% of students would be awarded an A* when their true grade is lower than this and, similarly, 1.6% of students would not be awarded A* when they deserve it. These values match as, by the definition we have used, the distribution of true grades equals the grade distribution on the full A Level. The dashed horizontal line demarks this level of over/under-grading and will be carried forward to the next chart.

For interest, we note that even for a full-length A Level, we can slightly improve classification accuracy with a slightly different grade boundary (the orange vertical line). Specifically, if we are interested in identifying the top 7.8% of students by true score (not observed score) then we should use a slightly harsher boundary (i.e., award fewer A*s). This is because, given that A*s are fairly rare, we need strong evidence to convince us that a student is most likely in this category rather than combining a slightly weaker true ability with a little luck with measurement error. Using this slightly harsher boundary would very slightly decrease the overall misclassification rate from 3.2% to 3.1%.

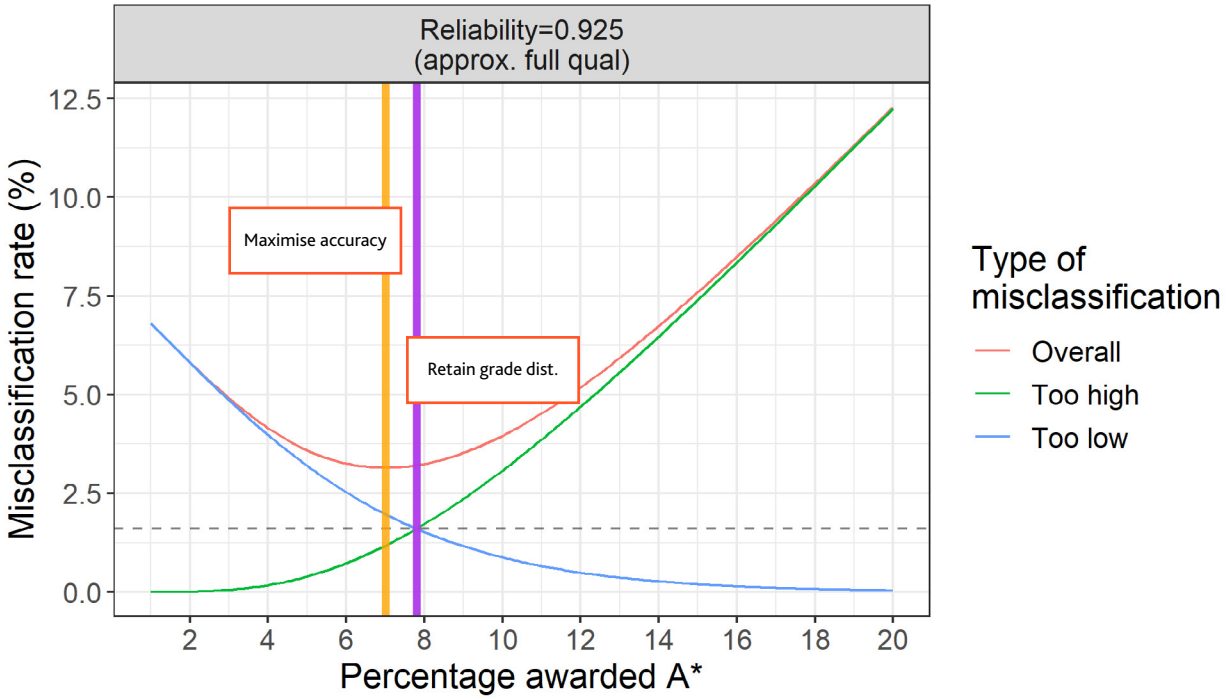


Figure 3: The relationship between percentage of students awarded A* and the misclassification rate at this grade for an overall assessment reliability of 0.925. The vertical lines denote pass rates associated with different awarding strategies.

Now we turn to our main topic of interest—the impact of our different strategies to setting grade boundaries as the reliability of the assessment drops. Figure 4 shows the relationship between the percentage of students awarded A* and the misclassification rate if the reliability of the assessment dropped to 0.72. The first thing to note is that, as we would expect, the overall misclassification rates rises. The pink vertical line indicates the results if we decide to continue to award an A* to the top 7.8% of candidates by their observed score despite the fact that we now have a less reliable assessment. Doing this results in the overall misclassification almost doubling from 3.2% to 6.3% with half of these students awarded a grade that is too low.

An alternative approach is to try to ensure that the proportion of students who are not awarded an A* when they deserve one remains at the target level (the dashed horizontal line, at 1.6%) that would be achieved by the full-length qualification. This goal is achieved by awarding A*s to 14.2% of candidates (the purple vertical line). In other words, in order to ensure that the reduction in reliability does not result in more students being under-graded than usual, we need to allow this many students to be awarded an A* in total. This point represents a logical maximum for the level of benefit of the doubt we might apply at A*. Going any further would mean even fewer students being under-graded than we would achieve in a normal year—not a rational response to a change in circumstances. Naturally, increasing the percentage of A*s we award to this extent will result in a big increase in the percentage of students who are over-graded. Specifically, this would mean that 8% of all students would be awarded an A* when they do not deserve one.

Finally, we consider the percentage of students we would award A* to if our goal was to maximise accuracy (i.e., minimise misclassification). This point is shown by the orange vertical line in Figure 4—reducing the percentage awarded A* to 4.7 per cent. This reduction in A* pass rate occurs because, given the unreliability of the assessment, we

need a lot of convincing that a student is genuinely among the top 7.8 per cent of true scores. In other words, we would need to see students even further towards the top of the observed score distribution.

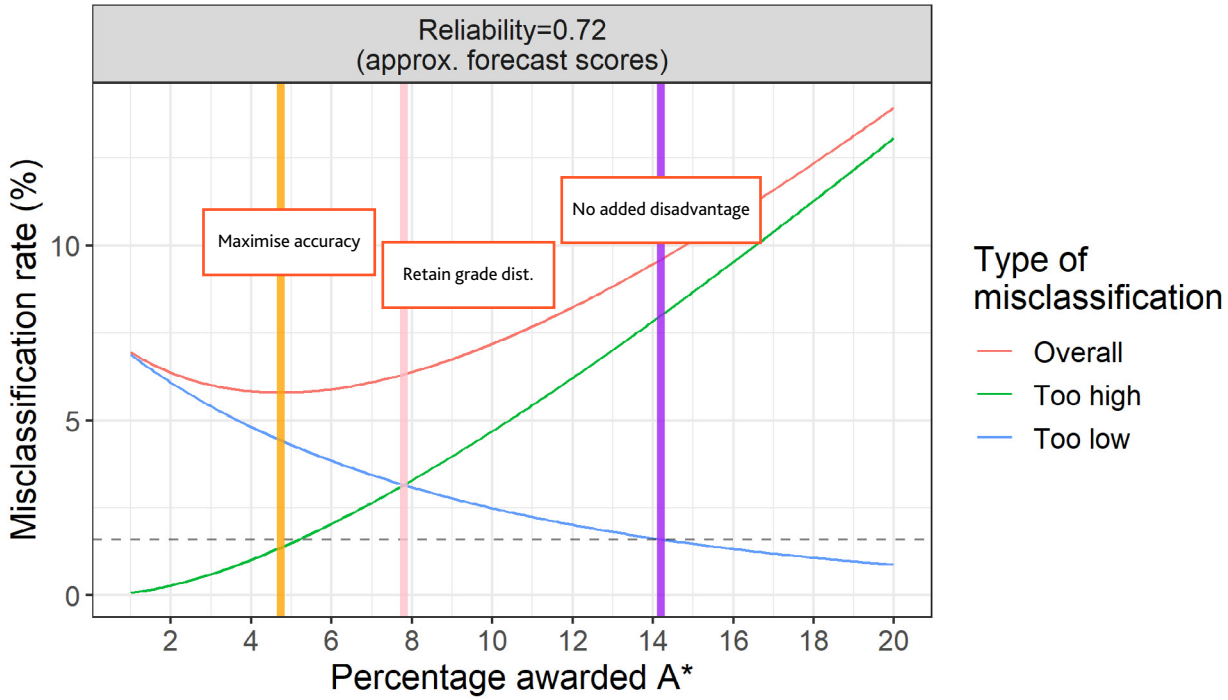


Figure 4: The relationship between the percentage of students awarded A* and the misclassification rate at this grade for an overall assessment reliability of 0.72. The vertical lines denote pass rates associated with different awarding strategies.

One way to look at Figure 4 is in terms of the balance between aggregate outcomes and individual risk. Assessment organisations are frequently accused of being too concerned with grade distributions at a national level and not enough on outcomes for individuals. With this in mind, it is worth noting that the horizontal axis in Figure 4 represents overall aggregate outcomes. Against this, the vertical axis represents the risk of individuals[7] being awarded a grade different from their true grade. The different strategies for choosing boundaries represent different choices regarding the balance of controlling aggregate outcomes versus managing individual risk. Retaining grade distributions places all of the emphasis on aggregate outcomes. The No Added Disadvantage strategy places all of the emphasis on ensuring the risk to individuals does not increase. It is worth noting that if our focus is on managing individual risk, while we would increase aggregate outcomes in a year where reliability is lower, we would subsequently reduce aggregate outcomes once reliability had risen back to normal levels (i.e., the situation in Figure 3).

Figure 5 shows the results of the same kinds of analysis focusing on grade E. Note that, based on the national A Level grade distribution in 2019,[8] for the purposes of calculations,

7 For example see https://www.theguardian.com/education/2020/aug/17/gavin-williamson-seeks-blame-ofqual-exams-debacle-a-level-gcse

8 See https://www.jcq.org.uk/wp-content/uploads/2019/08/A-Level-and-AS-Results-Summer-2019.pdf

97.6 per cent of students are defined to be at grade E or above. The bottom panel shows the results for a full-length qualification. Note that, since failing to achieve grade E is so rare, the overall misclassification rate at this grade is very low (just 1.2 per cent). Note also that, in contrast to grade A*, since true scores below grade E are rare, we need a lot of evidence to convince us that a candidate is not worth a grade E. As such, at this grade, classification accuracy is maximised by setting grade boundaries more leniently (the vertical orange line).

The top panel of Figure 5 shows the impact on misclassification if the assessment reliability drops to 0.72. As before, in order to ensure the number of students under-graded is the same as normal, the percentage of students awarded grade E and above should increase (the vertical purple line). However, what is different here from the analysis at grade A* is that, to maximise accuracy, the percentage awarded E or above should rise even further (the orange vertical line). This reflects the fact that, with an unreliable assessment, we would need a lot of convincing that a student was not worthy of at least a grade E.

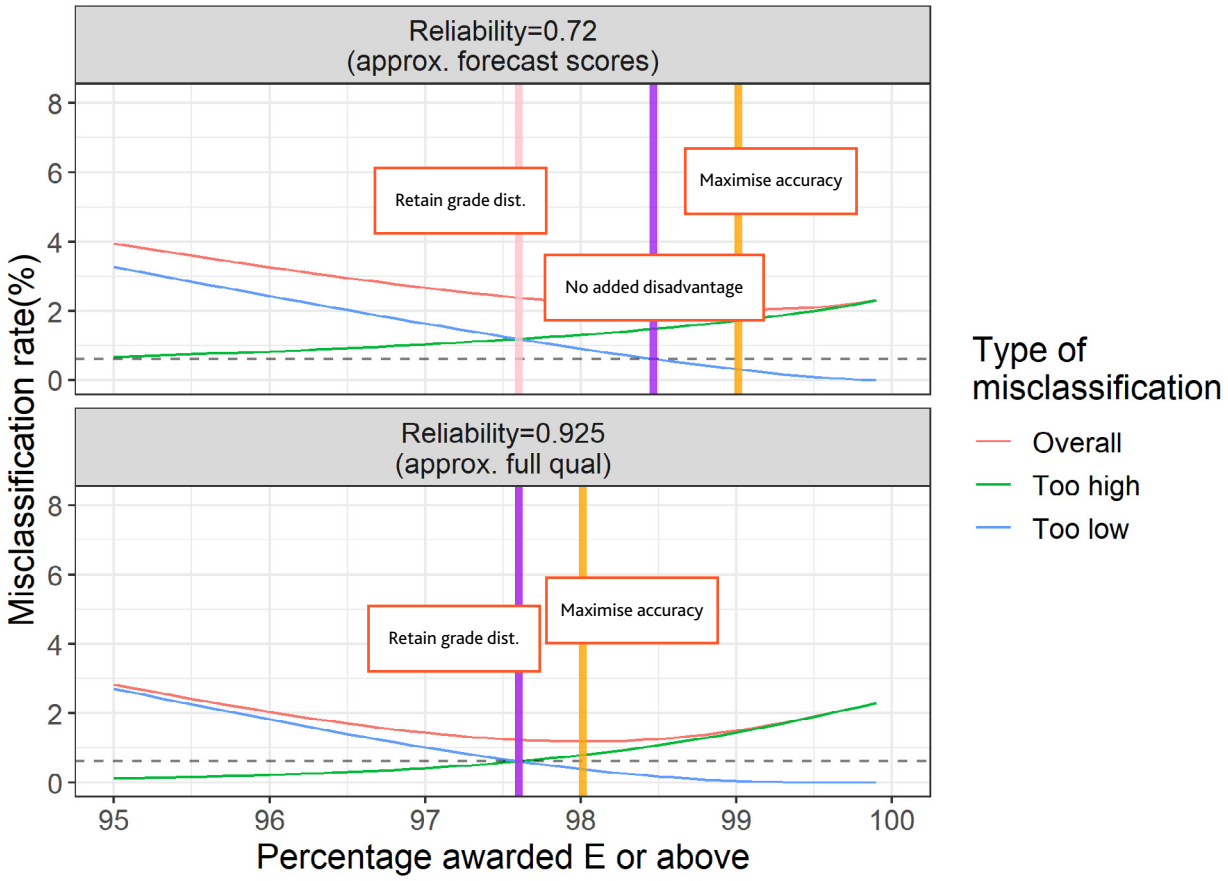


Figure 5: The relationship between percentage of students awarded E or above and the misclassification rate at this grade for overall assessment reliabilities of 0.72 and 0.925. The vertical lines denote pass rates associated with different awarding strategies.

The impact of different choices on grade distributions and misclassifications

Having illustrated our approach to thinking about the issue of using generosity to combat unreliability, this section presents a more complete set of results across all grades and across a greater range of assessment reliabilities. To begin with, we will present results relating to the strategy of ensuring that there is no increase in the percentage of students that are under-graded. After this, results relating to the strategy of maximising classification accuracy will be presented. In both cases we will compare results to the strategy of retaining the grade distribution regardless of the reliability of the test

Ensuring No Added Disadvantage

Table 1 shows how the No Added Disadvantage strategy affects the cumulative percentage awarded each grade as assessment reliability drops. The top row of the table (reliability=0.925) represents the assumed target grade distribution in a normal year. As can be seen, as the reliability of the assessment drops, so the required pass rate at each grade to avoid any increase in disadvantage goes up. For a small reduction in reliability (from 0.925 to 0.90), increases of around 1 percentage point at each grade are enough to mitigate the risks of decreased reliability. However, as reliabilities get lower, more drastic changes to the grade distribution are required.

Table 1: Percentage of students awarded each grade or above for different levels of reliability under the No Added Disadvantage strategy to setting grade boundaries.

Reliability	Cumulative % at each grade					
	A*	A	B	C	D	E
0.925	7.8%	25.5%	51.6%	75.8%	91.0%	97.6%
0.900	8.5%	26.8%	53.1%	76.8%	91.5%	97.8%
0.850	9.9%	29.4%	55.8%	78.7%	92.3%	98.0%
0.800	11.4%	32.0%	58.4%	80.4%	93.0%	98.2%
0.750	13.1%	34.6%	60.8%	81.8%	93.6%	98.4%
0.720	14.2%	36.2%	62.2%	82.6%	94.0%	98.5%

As described earlier, we estimate that the "reliability" of a teacher forecast approach to assessment is about 0.72. With this in mind, and taking this as our best guess of the reliability of CAGs, it is of interest to compare the grade distribution from the No Added Disadvantage approach with this level of reliability in Table 1 to the actual distribution of A Level CAGs awarded in summer 2020 (Table 2). What is interesting is that at grades A*, A and B there is a lot of similarity between a logical grade distribution based on the No Added Disadvantage strategy and the actual distribution of CAGs from teachers. For example, under the No Added Disadvantage strategy for a reliability of 0.72, we would increase the proportion of students awarded A and above from 25.5% to 36.2%. In reality, CAGs from teachers suggested that the proportion awarded A and above should rise

from 25.5% overall in 2019 to 37.6% in 2020. This may indicate that many teachers had a natural intuition for how confident they could be in their own estimates and applied a logical level of benefit of the doubt to help ensure that students were not disadvantaged relative to a normal year. Of course, the similarity in results may be purely coincidental. However, it does illustrate how major changes in grade distributions between years need not necessarily in themselves indicate inappropriate decisions. If our aim is to protect students from any adverse effects of added unreliability, a change in grade distributions is a logically justifiable result.

Table 2: Distribution of Centre Assessed Grades (CAGs) in summer 2020 (reproduced from Ofqual, 2020, Table 9.6, p.134)

	Cumulative % at each grade					
	A*	A	B	C	D	E
Centre Assessed Grades	13.9%	37.7%	64.9%	87.0%	96.4%	99.7%

Table 3 shows the overall misclassification rate of the No Added Disadvantage strategy across different levels of assessment reliability. Note that these misclassification rates are estimated for each grade separately. For example, the 7.1 per cent misclassification rate for grade A at a reliability of 0.925 means that we would expect this percentage of people to either have a true grade of A or above and be awarded a lower grade, or to have a true grade below A but to be awarded A or above. As one would expect, the lower the reliability of the assessment, the greater the percentage of students that are misclassified. The comparison of these misclassification rates to those we would get from a strategy of retaining the same grade distribution regardless of reliability (Table 4) is of more interest. As can be seen, for small reductions in reliability there is little difference between the overall accuracies of the two approaches. However, at grade A*, for the lowest reliabilities, the No Added Disadvantage approach does lead to substantially more misclassifications overall than the alternative strategy. Having said this, for grades C to E it actually leads to slightly fewer misclassifications than retaining the grade distribution (for an explanation see the earlier discussion surrounding Figure 5).

Table 3: Overall misclassification rates at each grade for the No Added Disadvantage strategy to setting grade boundaries for different levels of reliability.

Reliability	Misclassification rate at each grade					
	A*	A	B	C	D	E
0.925	3.2%	7.1%	8.8%	6.9%	3.6%	1.2%
0.900	3.9%	8.4%	10.3%	8.0%	4.1%	1.4%
0.850	5.3%	11.0%	13.1%	9.8%	4.9%	1.6%
0.800	6.8%	13.6%	15.6%	11.4%	5.6%	1.9%
0.750	8.5%	16.2%	18.0%	12.9%	6.2%	2.0%
0.720	9.6%	17.8%	19.4%	13.7%	6.5%	2.1%
0.700	10.4%	18.9%	20.4%	14.2%	6.7%	2.2%

Table 4: Overall misclassification rates at each grade if the same grade distributions are retained across different levels of reliability.

Reliability	Misclassification rate at each grade					
	A*	A	B	C	D	E
0.925	3.2%	7.1%	8.8%	6.9%	3.6%	1.2%
0.900	3.7%	8.2%	10.2%	8.0%	4.1%	1.4%
0.850	4.6%	10.2%	12.6%	9.9%	5.1%	1.7%
0.800	5.3%	11.8%	14.7%	11.5%	5.9%	2.0%
0.750	6.0%	13.3%	16.7%	13.0%	6.6%	2.3%
0.720	6.3%	14.2%	17.7%	13.8%	7.1%	2.4%
0.700	6.6%	14.8%	18.4%	14.3%	7.3%	2.5%

Maximising accuracy

Table 5 shows the cumulative percentage of students that would be awarded each grade if we adopted the strategy of maximising classification accuracy at each level of reliability. As can be seen, as the reliability of the assessment lowers, this strategy results in awarding fewer top grades (A and A*) but being more generous at lower grades (C, D and E). The strategy has little effect on the cumulative percentage awarded grade B or above. At the lowest grades (D and E) this strategy is even more generous than the No Added Disadvantage strategy (Table 1). However, comparison with Table 2 shows that the strategy would still not be as generous as CAGs awarded in summer 2020 at these grades.

Table 5: Percentage of students awarded each grade or above for different levels of reliability under the Maximise Accuracy strategy to setting grade boundaries.

Reliability	Cumulative % at each grade					
	A*	A	B	C	D	E
0.925	7.0%	24.7%	51.7%	76.7%	91.8%	98.0%
0.900	6.7%	24.4%	51.7%	77.0%	92.1%	98.1%
0.850	6.2%	23.7%	51.7%	77.6%	92.7%	98.4%
0.800	5.6%	23.1%	51.8%	78.3%	93.3%	98.6%
0.750	5.1%	22.3%	51.8%	79.0%	93.9%	98.9%
0.720	4.7%	21.9%	51.9%	79.5%	94.3%	99.0%
0.700	4.5%	21.6%	51.9%	79.9%	94.5%	99.1%

Table 6 shows the misclassification rates that would result from this strategy. Although (by design) these values are always lower than for the other two strategies we have explored (Tables 3 and 4), the difference with the strategy of retaining the same grade distribution is always very small. Specifically, for the range of reliabilities displayed here, the difference in misclassification rate is always within 1 percentage point.

Table 6: Overall misclassification rates at each grade for the Maximise Accuracy strategy to setting grade boundaries for different levels of reliability.

Reliability	Misclassification rate at each grade					
	A*	A	B	C	D	E
0.925	3.1%	7.1%	8.8%	6.9%	3.5%	1.2%
0.900	3.6%	8.2%	10.2%	8.0%	4.0%	1.4%
0.850	4.4%	10.1%	12.6%	9.8%	4.9%	1.6%
0.800	5.0%	11.7%	14.7%	11.3%	5.6%	1.8%
0.750	5.5%	13.1%	16.7%	12.7%	6.2%	2.0%
0.720	5.8%	13.9%	17.7%	13.5%	6.5%	2.0%
0.700	6.0%	14.4%	18.4%	14.0%	6.7%	2.1%

Conclusion

This article has considered how we might logically go about applying generosity to grade boundaries to address a situation where, for whatever reason, an assessment is less reliable than normal. We have seen that a rigid adherence to the idea that no-one should end up with a lower grade than they deserve is not practical. In fact, with any level of unreliability, and a sufficiently large number of candidates, such a goal could only be achieved by awarding the highest possible grade to every candidate. However, once we acknowledge that, even under normal circumstances, no assessment is perfectly reliable, we can make progress with this issue. In particular, we can determine the most appropriate grade distribution given an idea about how much reliability has decreased and how we wish to balance the competing desires to control aggregate outcomes and to manage the level of risk to individuals.

One strategy we might adopt is to maximise the overall accuracy of awarded grades. This will lead to different results than simply retaining the same grade distribution regardless of how far reliability has fallen. Specifically, it will make us more reluctant to award the very highest grades as it is difficult for an unreliable assessment method to verify that these are warranted. Conversely, under this same strategy, we would be more generous with awarding grades at the lower end as it is difficult for an unreliable assessment to verify that a candidate *does not* deserve to be awarded a grade of at least this level.

Another option is the strategy we have termed No Added Disadvantage. Under this option, which places the most emphasis on managing the risk to individuals, we set grade boundaries to ensure that the overall proportion of students awarded a lower grade than merited by their true ability does not increase as the reliability of assessment falls. This option requires generous awarding across all grades with the required level of generosity increasing to compensate for lower assessment reliabilities.

One interesting finding from this research comes from comparing the results from this strategy to the actual distribution of CAGs supplied by schools in summer 2020. By using data from forecast grades in the past, and by noting that CAGs were a form of forecast, we generated a data-driven estimate of the likely reliability of CAGs. Based on this, we found that, in part, the grade distribution in summer 2020 represented a perfectly reasonable application of benefit of the doubt by teachers to ensure that students were not negatively affected by the circumstances they found themselves in. The exception to this was at the lowest A Level grades where, based on our approach to calculations, teachers appeared even more generous than would be recommended by the principle of No Added Disadvantage.

The above paragraph has some implications for awarding once assessment returns to normal. Some would argue that the generous distribution of grades from 2020 should be carried forward into the future as not doing this is unfair to subsequent cohorts of students. However, this fails to recognise the possible role of benefit of the doubt in teachers assigning grades in 2020. As we have shown, at least at the higher grades, the 2020 grade distribution represents a perfectly sensible operation of benefit of doubt given how hard it is to know how students will perform in a future set of exams. However, if the same generous grade distribution is applied in future assessment series where there is far less doubt over performance, any intended benefit of the doubt by 2020 teachers is overridden. One way to think of this issue is as follows. It is possible that some teachers in

2020 awarded grades to students based on the highest grade they think they might have reasonably achieved. This helped protect individual students from being under-rewarded. However, had these same teachers known that the performance standards required for each grade were going to be lowered, they could have extended this same principle of benefit of the doubt further to other students—but did not. Lowering the required performance standards could re-introduce the possibility of students missing out on the grade they deserve relative to future cohorts of candidates due to the unreliability of the assessment procedure they were subject to in 2020.

To put this another way, where there is more doubt there is more benefit of the doubt. Applying this principle in a logical way means that more top grades should be awarded in a year of unusually unreliable assessment than in a normal year. Arguing that the grade distribution from 2020 should simply be carried forwards on grounds of fairness places too much emphasis on national statistics and not enough on individual students—something that assessment organisations are often accused of!

More generally, this article has discussed issues relating to awarding in the knowledge that assessments are not perfectly reliable. Any discussion of classification accuracy in qualifications is uncomfortable as it opens the door to sound and fury about "students being awarded the wrong grade". However, in reality, classification accuracy is just putting a quantitative value on an experience most exam takers are probably familiar with. Many people can probably remember hoping that certain topics or types of problem would be included in their exams and then being either pleased when they were or disappointed when they were not. Reliability research just attempts to put a number on the extent to which these types of familiar experiences have an impact upon grades. Similarly, the attempt in this article to quantify the reliability of teacher forecasts is not an attack on teachers—simply an effort to quantify how hard it is to know what grade a student will achieve months before they take an exam.

More importantly, it is only by acknowledging and quantifying likely levels of reliability that we can make specific recommendations around issues such as benefit of the doubt. The research presented in this article would not have been possible if we had tried to start from a point of view that in a normal year assessment is perfectly reliable.

Within any crisis, grade boundaries are an obvious element of assessment for people to focus on as, from a practical perspective, altering them is relatively easy. Nonetheless, it would clearly be naïve to think that all the problems created by a pandemic (or another unforeseen circumstance) can be solved simply by changing grade boundaries. After all, altering grade boundaries does not change anything about the fundamental measurements that have been made about students—it simply alters the way these measurements are labelled. With this in mind, although this article has focused upon grade boundaries, this is not intended to diminish the role of other elements of the assessment system in dealing with a crisis. Arrangements for appeals, additional assessment opportunities, advice to users of results (such as universities) and various other interventions can all be important parts of the ways in which we can alleviate the risk of individual students being disadvantaged.

References

Benton, T. (2014). Calculating the reliability of complex qualifications. *Research Matters: A Cambridge Assessment publication, 18,* 48–52. https://www.cambridgeassessment.org.uk/Images/465814-calculating-the-reliability-of-complex-qualifications.pdf

Benton, T. (2016). *Evidence for the reliability of coursework.* Paper presented at the 17th annual AEA Europe conference, Limassol, Cyprus, 3–5 November 2016. https://www.cambridgeassessment.org.uk/Images/346753-evidence-for-the-reliability-of-coursework.pdf

Bramley, T., & Dhawan, V. (2010). *Estimates of reliability of qualifications.* Ofqual/11/4826. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/578868/2011-03-16-estimates-of-reliability-of-qualifications.pdf

Gill, T., & Benton, T. (2015). *The accuracy of forecast grades for OCR A levels in June 2014.* Cambridge Assessment Statistics Report Series No.90. https://www.cambridgeassessment.org.uk/Images/241261-the-accuracy-of-forecast-grades-for-ocr-a-levels-in-june-2014.pdf

Hayes, M., & Pritchard, J. (2013). *Estimation of internal reliability.* Ofqual/13/5258. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/378056/2013-01-17-estimation-of-internal-reliability.pdf

He, Q. (2009). *Estimating the reliability of composite scores.* Ofqual/10/4703. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/578878/2010-02-01-composite-reliability.pdf.

JCQ. (2019). *GCE A Level & GCE AS Level Results Summer 2019*. https://www.jcq.org.uk/wp-content/uploads/2019/08/A-Level-and-AS-Results-Summer-2019.pdf

Mc Manus, I. C., Woolf, K., Harrison, D., Tiffin, P.A., Paton, L. W., Cheung, K. Y. F., & Smith, D. (2020). *Calculated grades, predicted grades, forecasted grades and actual A-level grades: Reliability, correlations and predictive validity in medical school applicants, undergraduates, and postgraduates in a time of COVID-19.* https://www.medrxiv.org/content/10.1101/2020.06.02.20116830v1.full.pdf

Ofqual. (2020). *Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: Interim report.* Ofqual/20/6656/1. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909368/6656-1_Awarding_GCSE__AS__A_level__advanced_extension_awards_and_extended_project_qualifications_in_summer_2020_-_interim_report.pdf

Shaw, S. D., & Nisbet, I. (2021). Attitudes to fair assessment in the light of COVID-19. *Research Matters: A Cambridge Assessment Publication, 31,* 6–21.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

Stewart, H., Proctor, K., & Weale, S. (2020, August 17). Gavin Williamson seeks to blame Ofqual for exams debacle. *The Guardian.* https://www.theguardian.com/education/2020/aug/17/gavin-williamson-seeks-blame-ofqual-exams-debacle-a-level-gcse

Wheadon, C., & Stockford, I. (2010). *Classification accuracy and consistency in GCSE and A Level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 To June 2009.* Ofqual/11/4823. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/578862/2011-03-16-aqa-classification-accuracy-and-consistency-in-gcse-and-a-levels.pdf.

Appendix: R code to create Figures 3 and 4

The code below shows how to create charts like those shown in Figures 3 and 4. The calculations in the remainder of the report were done in a similar way.

```
#code to create Figures 3 and 4
library(mvtnorm)
library(ggplot2)
#set main pass rate of interest (based on national A* percentage 2019)
pass_rate=7.8
#convert to boundary on standard normal distribution
boundary_main=qnorm(1-pass_rate/100)
#FIGURE 3
#set reliability of interest
reliability=0.925
#calculate error rates for a variety of boundaries
trial_pass_rates=seq(1,20,0.01)
trial_boundaries=qnorm(1-trial_pass_rates/100)
#correlation matrix between true and observed scores
sigma <- cbind(c(1,sqrt(reliability)),c(sqrt(reliability),1))
#overall percentage awarded a grade too low given their true score
error_rate_low=sapply(trial_boundaries
                      ,function(i) 100*pmvnorm(mean=c(0,0)
                                  ,sigma
                                  ,lower=c(boundary_main,-Inf)
                                  ,upper=c(Inf,i)))
#overall percentage awarded a grade too low given their true score
error_rate_high=sapply(trial_boundaries
                      ,function(i) 100*pmvnorm(mean=c(0,0)
                                  ,sigma
                                  ,lower=c(-Inf,i)
                                  ,upper=c(boundary_main,Inf)))
error_rate_tot=error_rate_low+error_rate_high
#identify target under-graded for Figure 4 (later)
targ_low=error_rate_low[trial_boundaries==boundary_main]
#make chart
min_error_x=trial_pass_rates[which.min(error_rate_tot)]
chartdat=data.frame(x=rep(trial_pass_rates,3)
                      ,y=c(error_rate_low,error_rate_high,error_rate_tot)
                      ,type=c(rep("Too low",1901)
                                  ,rep("Too high",1901)
```

```
                                        ,rep("Overall",1901)))
ggplot(data=chartdat,aes(x=x,y=y,col=type))+geom_line()+
        geom_hline(yintercept=targ_low,lty=2,alpha=0.5)+
        theme_bw()+scale_x_continuous(breaks=seq(0,20,2))+
        labs(x="Percentage awarded A*",y="Misclassification rate (%)"
                        ,col="Type of\nmisclassification")+
        theme(text=element_text(size=14))+
        geom_vline(xintercept=min_error_x,col="orange",size=1.5
,alpha=0.8)+
        geom_vline(xintercept=pass_rate,col="purple",size=1.5,alpha=0.8)+
        ggtitle("Reliability=0.925")
#FIGURE 4
#set reliability of interest
reliability=0.72
#correlation matrix between true and observed scores
sigma <- cbind(c(1,sqrt(reliability)),c(sqrt(reliability),1))
#overall percentage awarded a grade too low given their true score
error_rate_low=sapply(trial_boundaries
                        ,function(i) 100*pmvnorm(mean=c(0,0)
                                ,sigma
                                ,lower=c(boundary_main,-Inf)
                                ,upper=c(Inf,i)))
#overall percentage awarded a grade too low given their true score
error_rate_high=sapply(trial_boundaries
                        ,function(i) 100*pmvnorm(mean=c(0,0)
                                ,sigma
                                ,lower=c(-Inf,i)
                                ,upper=c(boundary_main,Inf)))
error_rate_tot=error_rate_low+error_rate_high
#make chart
min_error_x=trial_pass_rates[which.min(error_rate_tot)]
#find pass rate that gives closest to previous under-grading rate
NAD_x=trial_pass_rates[which.min(abs(error_rate_low-targ_low))]
chartdat=data.frame(x=rep(trial_pass_rates,3)
                        ,y=c(error_rate_low,error_rate_high,error_rate_tot)
                        ,type=c(rep("Too low",1901)
                                ,rep("Too high",1901)
                                ,rep("Overall",1901)))
ggplot(data=chartdat,aes(x=x,y=y,col=type))+geom_line()+
        geom_hline(yintercept=targ_low,lty=2,alpha=0.5)+
        theme_bw()+scale_x_continuous(breaks=seq(0,20,2))+
        labs(x="Percentage awarded A*",y="Misclassification rate (%)"
                        ,col="Type of\nmisclassification")+
        theme(text=element_text(size=14))+
        geom_vline(xintercept=min_error_x,col="orange"
,size=1.5,alpha=0.8)+
        geom_vline(xintercept=pass_rate,col="pink",size=1.5,alpha=0.8)+
        geom_vline(xintercept=NAD_x,col="purple",size=1.5,alpha=0.8)+
        ggtitle("Reliability=0.72"
```

A guide to what happened with Vocational and Technical Qualifications in summer 2020

Sarah Mattey OCR

In 2020, the COVID-19 pandemic and lockdown led to the closing of schools across the United Kingdom (UK), and to the cancellation of exams and assessments in England that were due to take place in the summer. The approach taken for GCSEs (General Certificate of Secondary Education) and A Levels was widely documented in the media. What Vocational and Technical Qualifications (VTQs) did was not the same as GCSEs and A Levels and was not covered in the media as widely. The different approach is because the structure and assessments of VTQs are very different from GCSEs and A Levels, while the approach also had to account for differences between different VTQs. While there may not have been as much coverage in the media for these qualifications, many learners sit VTQ assessments each year, where summer 2020 was due to be no different.

What are Vocational and Technical Qualifications?

VTQs cover a wide range of qualifications at different levels. In this article, we will concentrate on VTQs typically taken in schools and colleges that are usually sat by 14–18 year olds (from entry level to Level 3), similar to learners who normally take GCSEs and A Levels. Over 4 million of these VTQs were awarded in 2019/20 (Vocational and Technical Qualifications Landscape, n.d.) in a diverse range of subject groups such as: Arts, Media and Publishing; Construction, Planning and the Built Environment; and Health, Public Services and Care. Qualifications available include Cambridge Nationals, Cambridge Technicals and BTECs. Furthermore, analyses using data from the National Pupil Database from 2016/17 found that at Key Stage 4 (14–16 year olds) and Key Stage 5 (16–18 year olds) over 50 per cent of learners were taking at least some VTQs as part of their education course. Of learners taking Key Stage 5 Level 3 qualifications (the same level as A Levels), 24.9 per cent of learners were taking only vocational qualifications (Vidal Rodeiro & Vitello, 2020). Learners can use VTQs to enter the next stage of education, apprenticeships or employment (Department for Education, 2017).

Was the approach to VTQs the same as A Levels and GCSEs?

How General Qualifications (GQs) results were calculated did not differ between awarding organisations as the process was set out by the national regulator, the Office of Qualifications and Examinations Regulation (Ofqual, 2020d; Ofqual 2020e), so that all learners' results were calculated in the same way. This approach meant that a learner would not have a different result depending on the awarding organisation that provided it. However, the breadth and diversity of VTQs meant that such a generalised approach was unfeasible (Ofqual, 2020b). What works for one type of qualification offered by a

particular awarding organisation would not necessarily work for another, either due to the purpose, design and delivery of the qualification, or due to availability of assessment evidence (Ofqual, 2020b). In this vein, the regulator produced an Extraordinary Regulatory Framework (ERF) for VTQs, with the aim to ensure as much consistency in approach as possible (Ofqual, 2020c). This meant that the approach used for VTQs was based on the same underlying principles, even though different approaches were used for different qualifications.

The approach for VTQs included three options: calculate, adapt, and delay. The options that were used for a qualification depended on the type and purpose of the qualification.

Calculate

If a qualification was used for progression to further or higher education (sharing design features with GCSEs and A Levels), awarding organisations had to make sure a result was calculated in the absence of the learner being able to take assessments (Ofqual, 2020a), referred to here as a "calculated result". Likewise, awarding organisations needed to consider providing a calculated result first if the qualification was said to have a "mixed purpose" whose primary purpose was further or higher education. For these mixed purpose qualifications, awarding organisations were required by the regulator to consider how appropriate a calculated result would be, and if it was not appropriate, to then consider adapting the assessments (Ofqual, 2020a).

What evidence was used to calculate results?

For qualifications where awarding organisations needed to provide a calculated result in the absence of the learner taking the assessment, the method of calculation needed to be based on multiple sources of evidence potentially available to awarding organisations, as laid out in the ERF. The ERF provided information on what type of evidence was suitable to use in calculating results and the weight that should be given to certain types of evidence. Awarding organisations could then research, develop and implement a process to calculate results using such evidence, which was submitted to Ofqual and kept under review. While some of the sources of evidence were similar to those of GCSEs and A Levels, some sources of evidence were particular to VTQs, which made it likely that a different approach could be used.

Figure 1 summarises the evidence available to awarding organisations in producing calculated results that was provided in the ERF. The green boxes detail the type of evidence that could have been used while the orange boxes describe the weight such evidence should have been given and what that weight depended on. The purple boxes describe how that evidence could have been used, including whether it could have been used to calculate a result, or whether it could be used to check the calculated result, or provide additional information to add context to a calculated result or set of calculated results. One piece of evidence might have been used to calculate the whole result, or it might have been used in conjunction with other pieces of evidence that were combined to produce a calculated result.

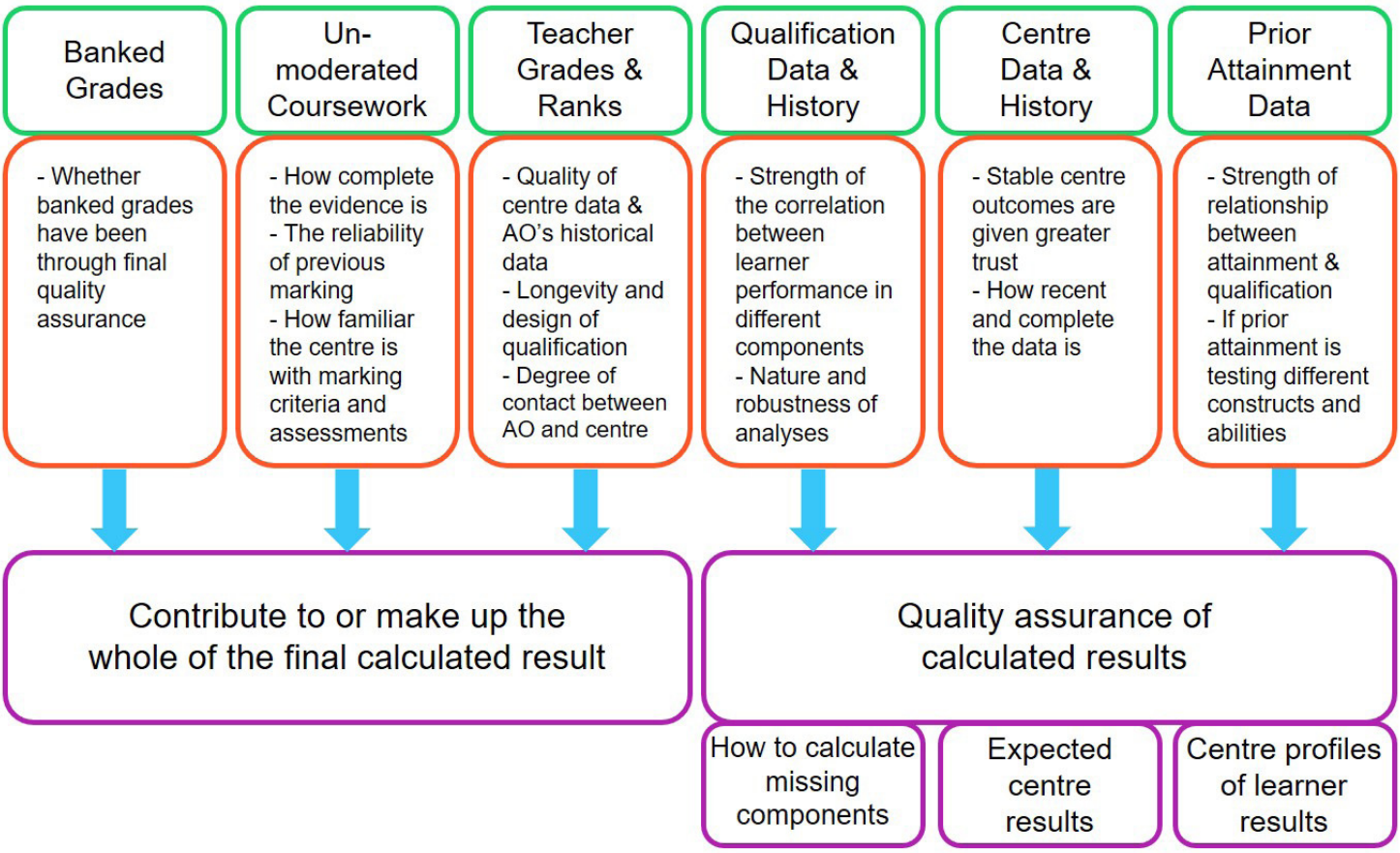


Figure 1: Illustration of the available evidence sources from Ofqual ERF that were available for awarding organisations (Ofqual, 2020c). Green boxes represent the type of evidence. Orange boxes represent what factors affect the weight to be placed on that evidence. Purple boxes represent how such evidence could have been used in producing calculated grades.

How did the evidence available for calculating results differ from A Levels and GCSEs?

There were some significant differences in the types of evidence that were available to VTQs compared to GCSEs and A Levels to calculate grades. Essentially, VTQs had more evidence of the work of individual candidates potentially available, though this could depend on the qualification or individual learners.

Firstly, many VTQs do not consist solely of terminal assessments or exams at the end of the course, known as a linear qualification. Instead, each component can be sat when learners are ready (for example, at scheduled sittings) throughout the period of study; this is known as a modular qualification. Modular VTQs mean that at any point in time, a number of learners will have already completed some assessment components before the end of the course and will have been awarded a mark that will eventually contribute to their final grade. These are referred to as banked components or grades. In this scenario, awarding organisations had some knowledge of these learners' performance in that subject and qualification, something that was not available to GQs due to their linear nature. Depending on how results were calculated and what was most appropriate for the qualification, awarding organisations may have used banked grades to estimate missing components that could then contribute to the qualification or contribute to calculating the qualification result in conjunction with other evidence. When available, banked grades were considered the strongest evidence available to awarding organisations for calculating

results and were given the highest degree of trust (Ofqual, 2020c). This is because banked component grades contribute to the final grade having been subject to standard (pre-COVID-19) quality assurance processes. However, not all learners have banked component grades, as this was dependent on the delivery model of the school or college, or learner. Therefore, it was likely that banked component grade evidence was available for some learners, but not others.

Further to evidence from banked grades, data about the functioning of the qualification could be used to determine whether it was reasonable to calculate results for missing components (where there was enough evidence) and for quality assurance (Ofqual, 2020c). In this instance, an important consideration about the weight to give banked grades was whether there was (historically) a strong correlation between outcomes on one component and other components within the qualification.

Another source of evidence available for some VTQs was any submitted internal assessment (Ofqual, 2020c), also known as un-moderated coursework or non-exam assessment (NEA). Some schools or colleges had submitted work but, because of COVID-19, the work had not been subject to standard quality assurance such as moderation. Without moderation of submitted coursework, the grade assigned by the centre may not have been the same as the grade that would have been given by the awarding organisation. To use this evidence, awarding organisations had to have a high degree of trust in the marking reliability of the centre and confidence that processes such as moderation would not change grades. This meant that many awarding organisations did not use this type of evidence to calculate results.

Did VTQs use some of the same types of evidence as A Levels and GCSEs?

VTQs could also use some of the same types of evidence that were available to GCSEs and A Levels for calculating grades. One of the primary sources was Centre Assessed Grades (CAGs) which provided the learners' likely performance estimated by schools and colleges (Ofqual, 2020c). CAGs from a college or school provided a judgement about the result that each learner was most likely to achieve. This grade might have been for the qualification as a whole or for individual components. As well as a grade, some awarding organisations may have asked for schools and colleges to also include a rank order of learners at each grade. For VTQs, the CAGs received from schools and colleges were subject to quality assurance by the awarding organisation. This may have been done statistically, for instance by comparing grades to additional evidence such as school achievement history, or by using a more qualitative approach by asking for additional evidence from schools and colleges. This additional evidence allowed awarding organisations to ensure that the schools and colleges had considered appropriate evidence when assigning the CAGs.

The ERF stated that when using such evidence, important factors to consider included the information schools and colleges had on learners (such as data from internal assessments), how modular the qualification was, how long the qualification had been available to learners, and the amount of contact the awarding organisation had with the school or college. In addition, the awarding organisation needed to consider whether the historical data it had could provide information on the accuracy of the grades (Ofqual, 2020c). How this type of evidence was used and the weight it carried may have been

different from GCSEs and A Levels, because schools, colleges and awarding organisations also potentially had additional types of current assessment evidence available such as unmoderated coursework or banked grades (Ofqual, 2020c).

As well as estimated performance from schools and colleges, other types of evidence that could provide additional information and context were also available to awarding organisations. While GQs used similar types of evidence, how this evidence was used may have differed for VTQs. Firstly, data on a school or college's historical performance on the qualification, for example how stable outcomes were and how well a school or college had performed, could be considered (Ofqual, 2020c). This could be used for quality assurance, framing expectations of centre level results, or for colleges and schools as a starting point when providing information to awarding organisations. Next, prior attainment data could be considered by awarding organisations. Such data could be how well a cohort did at Key Stage 2 (10–11 year olds) or the mean GCSE (15–16 year olds) scores of a cohort, as well as data on other relevant qualifications. For the prior attainment data, awarding organisations were asked to particularly consider how predictive prior attainment was of qualification outcomes. This is because VTQs have a more practical nature and can be designed to address different skills and constructs. Where appropriate, this evidence could have been used for looking at college and school profiles of learners' results or for quality assurances but could not be used for directly calculating learner results.

How was the evidence used to calculate results?

The variation in qualifications offered by different awarding organisations meant that unlike GCSEs and A Levels, there was also variation in how results were calculated, including whether they were calculated at the qualification or component level. The regulator considered and commented on how each qualification from the different awarding organisations was to be used to calculate results that met their criteria. This meant that different boards had to investigate and test approaches that were suitable for their qualifications (and deliverable within the timeframe). The outcome was that learners sitting different types of qualifications may have had their grades calculated in a different way, but in a way that was most appropriate for their qualification.

For example, if a learner did not have any grades banked, a result might have been calculated using CAGs. The awarding organisation may then have quality assured this by looking at the grade and any other grades in the context of a centre's results history, in order to ensure it was reasonable. This does not necessarily mean they would have changed the grade, as the context of other evidence and the weight they were given could have also been considered. An alternative for when a learner had no banked grades may have been to award a CAG after asking their school or college to provide evidence of how CAGs were awarded.

Another example is if a learner did have banked grades, a result may have been calculated solely based on this evidence, or it may have been combined with evidence from CAGs. The awarding organisation may then have looked at qualification data and history in terms of the relationship between different components to ensure it was fair to calculate grades for missing units using banked grades.

Due to the varied approaches used to calculate results for different qualifications, there is not a single summary of how these sources of evidence were used. Awarding organisations

produced their own communications detailing how results were calculated so that learners, schools and colleges, and the public could gain an understanding for individual qualifications.

Did the late decision to award GCSEs and A Levels based on CAGs affect VTQs?

In August 2020, after A Level results had been released, the method for awarding GCSEs and A Levels was revised. Statistical standardisation of CAGs to calculate results was removed as the default method, and learners were awarded either their CAG or their calculated result, whichever was higher (GOV.UK, 2020d). This decision for GQs had implications for some VTQ calculated results (GOV.UK, 2020c). It was important that VTQ learners would not be disadvantaged in comparison to GCSE and A Level learners. This meant that soon after the decision on GQs was made and depending on how grades were calculated and the purpose of qualifications (for example, whether learners took VTQs alongside GQs), some VTQ qualifications changed their approach to remain consistent with the GQ principle. Not all awarding organisations changed their approach, particularly where calculated results were based on banked grades, which remained a strong piece of evidence (GOV.UK, 2020a). That some qualifications changed their approach and others did not largely reflects the diversity in the initial approaches that was necessary for the different qualifications.

Adapt

For VTQs that were used to show occupational competence or provide a licence to practice, the awarding organisations needed to base results on the level of attainment demonstrated in an adapted assessment (Ofqual, 2020a). Awarding organisations also considered an adapted assessment as a starting point if the qualification was said to have a mixed purpose whose primary purpose was occupational competence (Ofqual, 2020a).

Adaptations could be made throughout the process, including the delivery, the assessment methods, the invigilation and quality assurance. Delivery of assessments could be adapted by ensuring tests could be taken online, or, for more practical scenarios previously face to face, assessments could be done remotely. Work experience or placements could also be adapted (for example by being shortened), or waived. Assessment methods could also be adapted, for example a practical simulation or discussion could be used instead of an observation or demonstration. Invigilation and quality assurances such as standardisation could also be adapted by moving online. An awarding organisation might have taken into consideration any relevant professional or sector body approach to how assessments were to be adapted, as particular professional or sector bodies may have been interested in maintaining consistency across qualifications awarded by different awarding organisations (Ofqual, 2020c).

Adapted assessments had to remain reliable, be able to maintain standards and be deliverable. It was particularly important to ensure none of the adaptations had an effect on the coverage of key areas of the construct that may have risked the validity of the qualification; if this could not be ensured then the assessments had to be delayed (Ofqual, 2020c). For professional qualifications or those that provide a licence to practice, it

was important to balance maintaining validity of the qualification while maximising the number of results awarded, though maintaining validity was prioritised (Ofqual, 2020c).

Delay

For all VTQs that could not be calculated or adapted without maintaining the validity of the qualification, assessments could be delayed, but this option should only have been considered where calculation or adaptation was not appropriate or possible (Ofqual, 2020a). Where assessments were cancelled, they should have been offered at later dates, such as by the autumn term (Ofqual, 2020a). Delaying assessments was a last resort option when calculate or adapt was not possible and ensured that as many learners as possible achieved a fair and valid result (Ofqual, 2020a).

What happened with VTQ assessments after summer 2020 and how will COVID-19 continue to affect VTQs?

With the nature of COVID-19, and its subsequent effects on schools and colleges, the awarding of VTQs in summer and autumn 2020 and 2021 has had to adapt. Because many VTQs are modular, awarding organisations were able to provide assessment opportunities after summer 2020. However, what assessments were possible to take will have varied for different schools and colleges.

The announcement in January that summer 2021 exams would be cancelled for GCSEs and A Levels also included that VTQ exams from February would also be cancelled. Exams in January could go ahead, though schools and colleges had a choice as to whether learners sat them or not (S. Lebus, personal communication, January 13, 2021). At the time of writing, how this decision will affect VTQ external and internal assessment in 2021 is in consultation (Department for Education & Ofqual, 2021). The effects of COVID-19 on the VTQ approaches highlight some of the differences between VTQs and GQs such as GCSEs and A Levels. The differences in qualification structures and assessments, as well as the diverse VTQ landscape have meant that they have not necessarily followed the same approach as GQs. However, while different methods were needed for different VTQs in 2020, the underlying principle of what evidence could be used and how it was applied was consistent. For 2021, while the differences between VTQ and GQ are still present in their structure and delivery, there are similarities in their purpose, as they are both used for progression to further education or employment. Because of this, awarding organisations and the regulator will look to ensure that, in comparison to learners taking GCSE and A Level, VTQ learners are not advantaged or disadvantaged.

Acknowledgements

I thank Breanne Chryst and Martina Kuvalja for discussions about the article. I also thank Frances Wilson and those who reviewed the article for their insights and constructive feedback.

References

Department for Education. (2017). *Technical and applied qualifications for 14 to 19 year olds. Key stage 4 and 16 to 18 performance tables from 2020: technical guidance for awarding organisations.*

Department for Education, & Ofqual. (2021). *Alternative arrangements for the award of VTQs and other general qualifications in 2021. Approach to awarding qualifications where exams do not take place* [Consultation]. Ofqual. https://www.gov.uk/government/consultations/consultation-on-alternative-arrangements-for-the-award-of-vtqs-and-other-general-qualifications-in-2021

GOV.UK. (2020a, August 17). *Grading of vocational and technical qualifications.* https://www.gov.uk/government/news/grading-of-vocational-and-technical-qualifications

GOV.UK. (2020b, August 17). *Statement from Roger Taylor, Chair, Ofqual.* https://www.gov.uk/government/news/statement-from-roger-taylor-chair-ofqual

GOV.UK. (2020c, August 18). *Ofqual statement on grading of VTQs, 18 August 2020.* https://www.gov.uk/government/news/ofqual-statement-on-grading-of-vtqs-18-august-2020

GOV.UK. (2020d, December 3). *Extra measures to support students ahead of next summer's exams.* https://www.gov.uk/government/news/extra-measures-to-support-students-ahead-of-next-summers-exams

Ofqual. (2020a). *Exceptional arrangements for assessment and grading in 2020. Consultation on the assessment and grading of vocational, technical and other general qualifications* [Consultation]. Ofqual. https://www.gov.uk/government/consultations/exceptional-arrangements-for-assessment-and-grading-in-2020

Ofqual. (2020b). *The awarding of vocational and technical qualifications and other general qualifications. Advice for teachers, trainers and tutors; learners, parents and carers; employers and professional bodies on vocational and technical qualifications, and general qualifications other than GCSEs, AS and A levels, Extended Project Qualifications and Advanced Extension Award in maths* [Guidance]. Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/904992/VTQ_Guidance_for_Centres_290720.pdf

Ofqual. (2020c). *Extraordinary regulatory framework Vocational and Technical Qualifications, COVID-19 Conditions and Requirements* [Regulations]. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/926985/VTQ_Covid-19_Conditions_and_Requirements_PUBLISHED_VERSION.pdf

Ofqual. (2020d). *Requirements for the calculation of results in summer 2020. GCE (AS/A level), GCSE, Extended Project and Advanced Extension Award Qualifications* [Requirements]. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/910614/6674_Requirements_for_the_calculation_of_results_in_summer_2020_inc._Annex_G.pdf

Ofqual. (2020e). *Summer 2020 grades for GCSE, AS and A level, Extended Project Qualification and Advanced Extension Award in maths. Guidance for teachers, students, parents and carers.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/908368/Summer_2020_grades_for_GCSE_AS_and_A_level_110820.pdf

Ofqual. (n.d.). *Vocational and Technical Qualifications Landscape.* https://analytics.ofqual.gov.uk/apps/VTQ/VTQLandscape/ (retrieved December 11, 2020).

Vidal Rodeiro, C. L., & Vitello, S. (2020). *Vocational Qualifications at Key Stage 4 and Key Stage 5: who takes them and how they fit into students' programmes of study.* Cambridge Assessment Research Report. Cambridge Assessment. https://www.cambridgeassessment.org.uk/Images/579348-vocational-qualifications-at-key-stage-4-and-key-stage-5-who-takes-them-and-how-they-fit-into-students-programmes-of-study.pdf

Early policy response to COVID-19 in education—A comparative case study of the UK countries

Melissa Mouthaan, Martin Johnson, Jackie Greatorex, Tori Coleman, Sinéad Fitzsimons[1]
Research Division

Introduction

In this article, we report on a series of analyses that we carried out during the early course of the pandemic. In early 2020, researchers in the Education and Curriculum Team in Cambridge Assessment Research Division started a project that we called Curriculum Watch. The aim of this project was to collate a literature and documents database of education and curriculum policies, research and analyses from across the four countries of the United Kingdom (UK).

Our work was inspired by the work of Raffe et al. (1999) who set out the positive benefits gained from comparing the policies of "the UK home nations"[2]. We anticipated that our Curriculum Watch database would provide a growing resource that could inform our understanding of education and curriculum issues beyond the UK. The literature base would also serve as an archive of events and our reflections on them. This is particularly useful as the presence of policy documents is sometimes short-lived, reflecting movements in policy decision-making.

As the pandemic hit the UK in February 2020, it affected education policy in ways that were difficult to foresee. It was also challenging at the time to make sense of the policy shifts that were occurring. This context presented us with an opportunity to use our literature base to make sense of the emerging educational picture.

Having a base of comparative policy and research literature allowed us to see how policy developments converged or diverged across the different countries of the UK, allowing us insight into some of the issues influencing education decision-making. We used our emerging literature data to conduct a multiple case study analysis to look more closely at the fast-changing policy contexts of England, Scotland, Wales and Northern Ireland as they responded to the challenges of the COVID-19 pandemic. Our observations covered the major pandemic-related education policy shifts that occurred in the six months from mid-March to mid-August 2020.

Case studies allow the exploration of issues in their context (Yin, 1981), for example allowing consideration of the historical or political factors that influence a policy shift, with the aim that the conclusions of the analysis can generalise beyond the specific

1 The work was carried out when the final author was a member of the Research Division.

2 A note on terminology: We use the term "nation" throughout this paper except when referring to the countries of the UK (since Northern Ireland is not a nation). Our use of the term "the home nations of the UK" is a direct reference to the terminology used by other scholars.

cases studied. By extending the method to a "multiple-case study" it is possible to gain insight into the complex and nuanced particularities of single cases through placing them alongside a group of broadly similar cases (Heale & Twycross, 2018, p.7).

In the next section, we will outline some of the methodological considerations that underpin Curriculum Watch. We go on to analyse some of the key areas of UK policy formation and content (in relation to curriculum, pedagogy and assessment) that we observed during the first six months of the unfolding pandemic. We then reflect on our analysis, considering the impact of the methods that we have chosen.

The benefits of comparative study methods

It is argued that there is an inevitable degree of implicit comparison needed to make sense of anything (Laming, 2004; Wilson, 2011). Cross-national study is perhaps an obvious methodological choice when trying to understand the policy conditions of any one nation. By juxtaposing observations, there is the potential for similarities and differences to be identified. It is also important to consider the degree of closeness of case studies when using this method. Bryman (2016) for instance argues that selecting case studies that are substantially similar in many aspects is appropriate, where differences found between the cases in the analysis are less likely to be attributable to inherent differences between the cases at the outset (p.68). A methodology of "most similar comparisons" (Paterson & Ianelli, 2007, p.332) allows careful study of changes in parallel systems and enables policy learning (Raffe, 2005). This is because the closeness of selected cases reduces the number of variables that need to be explained through analysis.

Multiple case studies have been favoured in social research as, it is argued, insights from multiple cases allow for theory-building to a greater extent than a single case approach (Dyer & Wilkins, 1991; Eisenhardt, 1989). The closeness of contexts also means that it is possible to construct a "policy laboratory" in which "different solutions to common problems can be put into practice simultaneously" (Paun et al., 2016, p.11), offering the opportunity for evidence exchange and policy learning across the UK (The Institute for Government, 2019). This closeness also allows each country to see what is possible in a context that is largely similar to their own (Croxford & Raffe, 2014).

As well as generating insight into how governance works across the different parts of the UK (Simkins, 2014), comparative study downplays the importance of the social and contextual particularities of the systems that lie within national boundaries, and heightens a focus on the structural similarities and interdependences of neighbouring systems (Raffe et al., 1999). Although this contextual downplaying can be seen by some to be a weakness of the method, it is considered by others as a way of undermining "container thinking" (Lidher et al., 2020, p.6). This type of thinking can be problematic as it can lead systems to over-focus on internal policy shifts without acknowledging the influence on their policy environment of factors stemming from beyond their own borders.

One of the premises of cross-national study is that borders between different nations are porous and that there is interdependence between nations. The question is whether this same premise can be applied to the countries of the UK, with an assumption that the UK is "a system" containing a set of similar, regional cases. This question is especially pertinent

since the different countries of the UK have diverse historical and cultural characteristics. This has led some to argue that the UK policy context is more accurately described as "a disunited Kingdom" (Donnelly & Osborne, 2005, p.148), with this disunity increasing with the devolution of administrative and executive responsibilities for education policies in Scotland, Wales and Northern Ireland since the 1990s.

UK policy: system or systems?

Some commentators point to significant features of overlap between the constituent countries of the UK, which means that they have more in common with each other than either of them does with any other national system. Drawing on Richardson's (1982) original analogy of different national policy "styles", Cairney (2013) argues that there is an "impressive degree of policy style convergence" between the UK countries (p.8). Within this, there are still specific national traits: the Scottish Government's implementation style involves closer engagement with local authorities in a "bottom-up" approach that differs from the other UK countries' top-down control (Cairney, 2013, p.8; see also Keating, 2010). The qualifications systems of England and Wales have been closely linked since 1917, while the Northern Irish qualifications system has drawn closer to that of England and Wales since its political partition from the rest of the island of Ireland in 1920. England, Wales and Northern Ireland have based their curricula on the same statutory guidance since 1988 (although the Welsh Government is currently in the process of implementing a new Curriculum for Wales that is not based on this guidance). Raffe et al. (1999) point to the overwhelming similarities between the social contexts and social relations across the four countries of the UK in comparison with those of other nations. They note that there are similar levels of social mobility and class inequality across the four countries, and that gender differences across the four countries are more alike when compared with those of many other nations. Finally, Paterson and Ianelli (2007) and Raffe et al. (1999) observe that the systems are economically interdependent, with all four education systems interacting with an economy which is integrated and organised at a UK level. As a result, we would expect policy influences to drift across the borders of the UK countries.

Another feature of our comparative methods approach is that it has a longitudinal dimension, attending to the way that policy changes over time. This policy change is also situated in the context of the COVID-19 pandemic, which can be characterised as an unfolding crisis. To support our analysis, we also look to literature that considers policy decision-making in times of crisis.

Crisis response literature

Our study makes use of a unique and unusual opportunity to examine simultaneous government responses to a common challenge, and to study the factors and conditions that drove decision-making processes. A common limitation of comparative policy scholarship is that issues unfold over a long period of time, and key aspects of decision-making are often difficult to identify or disentangle as a result (Capano et al., 2020; Grødem & Hippe, 2019). However, the COVID-19 pandemic constitutes a common and transboundary problem to governing elites that prompted a response in a specific time frame. It is evident that a body of literature analysing policy communities' responses

in different case study countries has already started to take shape.[3] This article adds to the emerging literature in this field. In addition, while crisis management is dealt with in many social science disciplines, including international relations, business studies, public administration, and communication studies, there is a general gap in research on collective "lesson learning" in the aftermath of crisis (Boin et al., 2018, p.33; Broekema, 2016; Hart & Kuipers, 2018). Simply cataloguing the policy measures adopted to date generates a database of raw data that can be used for such policy learning. Going beyond cataloguing measures, further and systematic analysis can then highlight interrelationships or important variables such as timing and sequencing (Capano et al., 2020; Howlett, 2019).

There are, moreover, key insights that we can draw from the crisis management literature that are relevant for understanding education policy responses in the UK countries. Policy literature and organisational studies examining institutional responses show, for example, that crises demand coordination at multiple levels within a network of organisations that may not be used to working with each other (Boin et al., 2018). Public bureaucracies, such as ministries, can often be seen to adapt poorly to crisis circumstances, given that routine business in a government body or public agency may require a radically different set of operating principles than crisis management, which requires flexibility and improvisation (Boin et al., 2018). The pandemic, as an external shock event, raises questions about how national policy response was constructed by governments, both cognitively and ideologically, and whether national policy styles (Howlett & Tosun, 2019) are discernible in the formulation of this response. How effectively governments manage and deploy policy narratives in times of crisis is also a factor that can vary widely between countries (Mintrom & O'Connor, 2020), even when comparing countries with similar policy regimes. In addition, the literature suggests examining not only variety in the "composition of the policy mix", but also variety in the timing of policy adoption, as well as the "intensity" or "stringency" with which various tools are deployed (Capano et al., 2020, p.297; Knill et al., 2012).

Cross-national study method

A principal aim of our Curriculum Watch initiative was to create a mechanism for locating and bringing together curriculum documentary evidence relating to the four countries of the UK. To do this, each researcher in the Education and Curriculum team took responsibility for gathering relevant documentary sources for one of the countries of the UK. This longitudinal source-gathering exercise allowed us to construct a picture of curricular policy change as it emerged in each country.[4] The second aim was to gather commentary and discourse related to curriculum policy change so that we could better understand any of the broader socioeconomic issues relating to such initiatives, but also

3 See for example: Masri and Sabzalieva (2020) on Higher Education responses in Canada; Sibieta and Cottell (2020) on education policy responses across the UK nations to the pandemic; and Mintrom and O'Connor's (2020) analysis of variation in policy response at the US state level.

4 For example, there is currently a great deal of interest in Scotland's Curriculum for Excellence. It was implemented in 2010 and the Organisation for Economic Co-operation and Development (OECD) is commissioned to review it in 2021.

stakeholder responses. For this reason, we included a range of documentary sources such as journal papers, conference outputs, news articles and blog posts.

We developed a database to bring together and organise our literature. This involved establishing a coding structure for documents added to the database. These codes (which include tags such as "Assessment", "COVID", "Curriculum structure/organisation", "Deprivation/Disadvantage/Vulnerable learners", etc.) allowed us to retrieve documents according to theme if we had a specific analysis that we wanted to pursue. We also constructed timelines for each of the countries. The timelines included the significant dates for educational and curricular events or developments in those countries. This process helped us to be aware of significant pieces of information that would be important to us when interpreting or analysing documents from a specific country.

In March 2020, as the impact of the pandemic on UK education policy was becoming more apparent, we decided to observe, document and monitor the emerging policy picture across a number of themes (policy in the first three months of the pandemic; changes to assessment; learning access, resources and assessment guidance; and curriculum choices).

This article builds on the outcomes of these observations (which were written in the moment of the shifts that they reported)—taking a step back in perspective. This step back allows us to carry out systematic analysis that can highlight interrelationships or important variables that may have not been visible at the time of the original data collection.

So far we have explored the debate around the extent to which the UK has one or many policy systems, and the potential of comparative study methods for gaining insights into the taken for granted (or "container") thinking that can potentially limit the understandings of how a policy system works. We have also suggested (drawing on the crisis response literature in public administration and organisational sociology) that an analysis of the unfolding policy story that occurred during the pandemic can provide an evidence basis for later "lesson learning". In the next section we describe the policy picture that emerged in the first six months of the pandemic across the four countries of the UK. This analysis looks specifically at curriculum guidance, pedagogy and assessment, and proceeds as follows. We first examine key similarities and differences in policy formation in the four countries in education response, including the involvement of different stakeholders in the policy process. We then focus on differences and overlaps of policy content in curriculum, pedagogy and assessment.

Analysis

Policy formation

The closing of schools triggered debate early on in all four countries on: the challenges of alternative assessment arrangements in light of the common decision to cancel the 2020 summer examinations; challenges around delivering the full curriculum; and how curriculum considerations would be integrated into subsequent recovery plans. In their responses, national governments of the UK countries demonstrated some differences in their policy formation approach and the level of guidance given to schools and local authorities. England opted for a more detailed and prescriptive approach to curriculum

guidance, publishing updated guidelines periodically and with individual guidance tailored to different Key Stages (Department for Education for England, 2020c). Wales and Northern Ireland, however, issued guidance that was general and less prescriptive, advising schools to be "responsive to changing circumstances" (Welsh Government, 2020a, p.2) and to adapt to the current context while "continuing to deliver the spirit of the statutory Northern Ireland curriculum" (Department of Education for Northern Ireland, 2020a, p.2). The detailed advice on the curriculum at secondary level that is evident in the Department for Education's (DfE) guidance for England is absent from national guidance issued in the other three countries. In addition, not all UK countries issued updated national guidance in view of rapidly changing contexts. As such, while curriculum guidance for England was issued after it became clear there would be a return to face-to-face learning, Northern Ireland's curriculum guidance (published end of June) was not updated in light of this development.

The Scottish approach of devolving key aspects of curriculum decision-making and curriculum preparation to local authorities also makes the Scottish response distinct from the other countries. As with Wales and Northern Ireland, the Scottish Government issued broad guidelines for schools regarding the curriculum. Yet, curriculum preparation in the recovery has been largely entrusted to local actors including Early Learning Centres, schools and partnership colleges (Scottish Government, 2020e). The role of local authorities therefore makes for an interesting comparison. The Scottish policy approach ensures local authorities and schools have a key role in policy enactment, and as a result, local Scottish authorities have provided detailed curriculum implementation support with guidance issued on planning and timetabling, but also support mechanisms for teachers. This is evident in for example Glasgow City Council's (2020) Recovery, Resilience and Reconnection framework. This arrangement of formal devolvement to local authorities is absent in England and Wales. However, councils in England and Wales have nonetheless been proactive in issuing their own guidance to schools, such as by outlining possible strategies to address lost learning and student wellbeing. For instance, Coventry City Council (2020) hosted a digital platform of information to support the mental health of learners, updating this as government advice became available.

Evidence from another UK study provides an early indication that pandemic responses that closely involve local authorities may see more success in terms of policy outcomes: Scottish and Welsh responses on the provision of free school meals proved more effective as they were delivered by local authorities, using existing infrastructure and demonstrating flexibility to families' needs (EPI Report: Sibieta & Cottell, 2020).

The issue of stakeholder engagement, and more specifically the transparency of such engagement, differs across the four countries. In the public discourse around the formation of policy on school return in Scotland, Wales and Northern Ireland, there is evidence of the range of voices who were consulted in this process. For example, in Scotland the C-19 Education Recovery Group included teaching union and local authority representatives, with the minutes of these meetings being published on the Scottish Government's (2020a) website. In Wales, the National Education Union (NEU) outlined that it was having regular meetings with the Welsh Government around the arrangements for school reopening, while in Northern Ireland, these policy decisions were "co-designed by the Department of Education, school leaders and key partners comprising Managing Authorities, trade unions and sectoral support bodies" (Department of Education for

Northern Ireland, 2020b). All of these cases contrast with the situation in England where the discourse was centred on headteachers and teaching unions "telling the government for weeks that the reopening plan was unworkable" (Weale, 2020), with requests from the Trades Union Congress (TUC) to include unions in the taskforce to plan return to school (TUC, 2020). This is also evident in the decision in England to hold an autumn exam series: while this decision was made following a consultation between the Office of Qualifications and Examinations Regulation (Ofqual) and stakeholders in April–May, schools and colleges appear to have had their concerns about available support for exam delivery amid schools reopening, and associated costs, largely sidelined.

A possible consequence of differences in consultation approaches is that it might affect the extent of professional engagement with policy content, and more specifically with policy change. On the one hand, we note a lack of clarity on key issues such as arrangements for 2021 examination sessions in the analysed time frame, as definitive decisions about future assessment were not forthcoming. In other instances, policymakers' responsiveness has meant that policy change was at times rapid, and sometimes resulted in a reversal of previous positions. The policy around opening schools for face-to-face education around the summer of 2020 was one such case. In England, where stakeholder engagement with policymaking appeared most opaque, there was a great deal of resistance to this policy from teacher unions, leading to a government U-turn around when primary school pupils would return (Weale, 2020). In contrast, when the Welsh Government announced that all students should return to school (not just specific year groups), there was negotiation with local authorities to determine term dates and how schools would re-open, which was welcomed by the National Association of Head Teachers (NAHT, 2020) union.

Policy content

In this section we look at some of the key curriculum, pedagogy and assessment content areas that were covered in policy decisions in the first six months of the pandemic across the countries of the UK.

Curriculum

A key issue that the four countries have grappled with in the wake of the pandemic is the question of curriculum adaptation, and finding the balance between core learning (e.g., a focus on reading, vocabulary, writing and mathematics) and breadth of learning in difficult and uncertain circumstances. Maintaining the full breadth of the curriculum, where possible, emerges as an important consideration, particularly in the national curriculum guidelines for secondary level published in England. England's guidelines show that flexibility is encouraged up to Key Stage 3 (ages 11–14), but schools are advised to discourage their pupils from dropping subjects altogether (DfE for England, 2020e). The requirements of qualification specifications for Key Stages 4 (ages 14–15) and 5 (ages 16–18) mean that they lack the flexibility allowed in earlier Key Stages. The rationale underpinning this guidance is that maintaining a broad and balanced curriculum ensures choices for further study and employment are not reduced, and that learners are taught a wide range of subjects (DfE for England, 2020e). In comparison, in Wales, Northern Ireland and Scotland, there has been more allowance for flexibility, and guidelines are less prescriptive. Scotland's guidelines have proposed a minimalist, core-learning

approach (Scottish Government, 2020c), and beyond this general indication, the Scottish Government encourages curriculum planning at the local level. In Northern Ireland, the curriculum is in any case ordinarily non-prescriptive and flexible. The decision in Wales to temporarily suspend basic curriculum requirements in June (Welsh Government, 2020b) is also an indication that Wales is prepared to adopt a flexible approach to the curriculum in its policy response.

The ongoing debate on how to balance "catch up" learning in the curriculum with student wellbeing is one that has drawn wider stakeholder debate in each of the four countries. The Welsh Government's plan for managing student wellbeing is particularly comprehensive, and is likely a reflection of the Government's commitment to emphasise health and wellbeing in the curriculum prior to the pandemic (see Public Health Wales, 2020). In Scotland, there is a similar emphasis on health and wellbeing in the published national guidance. Pupils in Scotland have experienced the shortest period of school closure in the UK (11 weeks, compared to 12–14 weeks in other UK countries). This could be an explanation of why guidance provided by some Scottish councils (see e.g., Glasgow City Council, 2020) favours short-term solutions to facilitate pupils' return to face-to-face learning: that is, focusing on student wellbeing in the first two weeks following the return to school to help pupils "reconnect".

Pedagogy

Two areas of pedagogy that were heavily discussed during this stage of the pandemic response were around blended learning arrangements (including access to technology and concerns around equity), and around the safety of pupils' physical return to schools.

There was a common concern across all of the countries with regard to ensuring disadvantaged learners were not adversely affected by the physical closure of schools to most learners at the start of the lockdown period in March 2020 (DfE for England, 2020a; Scottish Government, 2020b; Welsh Parliament Research Service, 2020; NI Direct, 2020). Despite their shared concern, there was evidence that the different countries approached the issue differently. For example, some targeted digital resource support to all disadvantaged learners (Scottish Government, 2020d), while others initially targeted this to specific age groups (DfE for England, 2020b).

When looking at the conditions around teaching in post-lockdown classrooms, there were clear differences in the measures taken by some of the countries around monitoring (and informing) classroom practices. In Scotland, inspections were put on hold (TES, 2020), while in Wales, inspections were suspended to enable inspectors to support the curriculum through engagement visits (Estyn, 2020). In Northern Ireland, inspections were suspended but there would be visits to check on compliance with COVID-19 safety rules (but these would not be shared with parents) (O'Brien, 2020). These positions contrast with the approach in England where the Office for Standards in Education, Children's Services and Skills (Ofsted) announced it would begin school visits to establish how they were "getting back up to speed", with normal inspections (which seek to grade schools and make judgements) planned to return in January 2021 (Professional Association for Childcare and Early Years, 2020). This suggests differences in emphasis, with England appearing to signal a more immediate return to external performance monitoring (even if this is not the intended message) compared with the teacher support focus of inspection visits for teachers in Wales and Northern Ireland.

Finally, arrangements for planned off-site education in international settings were equally affected across the four countries. Three policy announcements by the UK Government were directly referenced by the advice conveyed by some of the devolved administrations. Initially, the UK Government advised schools and colleges to cancel overseas trips in early March (UK Government, 2020a). The Foreign and Commonwealth Office (FCO) then advised British people to return immediately to the UK less than two weeks later (UK Government, 2020b). This advice was subsequently superseded by FCO advice around travel corridors to specific countries in July (UK Government, 2020c). This advice had an undoubted effect on planning for school trips across the countries of the UK. Advice from both the Scottish and Welsh Governments reiterated the FCO advice and advised against international educational visits (Scottish Government, 2020f; Welsh Government, 2020c).

Assessment

One of the key elements of the education response debate during the pandemic has been the question of assessment. As all UK countries cancelled the 2020 summer examination series, attention quickly shifted to arrangements for awarding student grades by August. This took on a broadly common formula of a two-step process of grade calculation: in the first instance, centres and teachers would decide the grade that each learner was likely to have achieved had they sat their exams. This grade would draw on teachers' judgements, and evidence such as prior work and mock exam results. This data would subsequently be used by the awarding body, together with a rank order list of students provided by schools, and a centre's historical performance, to calculate a final grade.

While these alternative awarding arrangements emerged as broadly similar, they also had their own specificities. For instance, England and Wales could be seen to have adopted a similar approach, but with some differences in the type of data used, as the Welsh and Northern Ireland grade calculation models used different prior attainment data to England (Qualifications Wales, 2020).

The scope for appealing grades was an area where countries adopted a different approach. Students in the different countries could appeal grades through their centre on specific grounds, for example if it was suspected that either the centre or the awarding body had made a technical error. Scotland additionally set up a priority review system for appeals from candidates with a conditional university or college offer (Scottish Qualifications Authority, 2020). It is evident that the scope for appealing Centre Assessed Grades (CAGs) in England was more limited from the outset in comparison to the other three countries: this is because it was determined early on that a full autumn exam series would be available as an option for students dissatisfied with their calculated grades (Ofqual, 2020).

With regard to formative assessment, national guidance issued in England recommends the use of regular formative assessment, such as quizzes and pupil observation, as a tool in classrooms to determine new "starting points" in education, and to ascertain the scale of lost learning (DfE for England, 2020c). Scotland's national guidance clarifies expectations that the Curriculum for Excellence will continue to apply, and also recommends teachers to use informal assessment to collect evidence on learners' progress, but to consider that formal tests may not be the most appropriate approach during the early recovery phase (Scottish Government, 2020c). This advice is broadly mirrored in Northern Ireland where

the Department of Education (2020c) advised that teachers should be using formative assessment as a check on pupils' learning, with greater discretion for schools on what to include in school reports when statutory assessment arrangements were suspended.

Analytical reflection

In this study, we used a comparative cross-country study method to gain insights that might not otherwise be apparent from isolated, individual case study analysis (Raffe et al., 1999). Looking to the methodology of "most similar comparisons" (Paterson & Ianelli, 2007, p.332), we recognise the closeness between the national systems of the UK (e.g., based on their shared geographical boundaries, population movement, historical connections, cultural and linguistic links). This closeness means that there is less noise that can account for differences in the comparison of these systems (compared with more dissimilar systems) and enables policy learning (Raffe, 2005, p.2).

Our chosen methodology gives us a clear rationale for systematically gathering curriculum policies and allied documents into a framework. The methodology allowed us to focus on co-occurring issues (e.g., around responses to the COVID-19 pandemic) and to consider convergence and divergence in policy responses across the countries, reasons for these patterns, and whether there are issues of interdependence between the countries' systems. The narrow time frame of our analysis, covering a period of a few months, also allowed us to focus in detail on key aspects of decision-making that would be more difficult to unpack over a longer time frame (see Capano et al., 2020). In the remainder of this section, we detail this article's key thematic findings and conclude with some final methodological reflections.

Divergence in governance styles—but common concerns

As the pandemic has unfolded it has required policymakers to choose between (i) fast response and (ii) delayed response, and in each case balancing sometimes conflicting issues such as protecting public health, ensuring educational access and supporting economic growth. We noted that there is evidence of both rapid decision-making and delayed decision-making. Delays in decision-making can be strategic, as governing bodies await further clarity on the fluid public health situation. We found that differences in consultation styles had an impact on how rapid policy shifts were received by stakeholders.

Cairney (2013) argues that there is an "impressive degree of policy style convergence" (p.8) across the constituent countries of the UK, but our analysis points to several areas of divergence. There are differences in the degree of transparency in stakeholder engagement in policy formation when looking at England compared with the other countries. It is possible that one consequence of the opaqueness of the involvement of multiple voices in policy formation is that it has an impact on reactions to subsequent policy shifts. The reactions of teachers and unions to the alterations to school re-opening dates were largely more negative in England compared with reactions in Wales.

Building on this observation, and aligning with Howlett and Tosun's (2019) assertion that "national policy styles" (p.4) are often evident in policy formulation, our analysis suggests that there are stylistic differences in policy formation and enactment across

the UK countries, with England appearing to exhibit a more "top down" or centralised approach than the other countries. This point is reinforced by Machin et al. (2013) who note that local authorities are more involved in the implementation of education policy in Scotland and Wales compared with England. This difference appears also to be reflected in the differences in the level of prescription in educational guidance, and in the approach to school inspections in England compared with the other countries. These observations lead to tangible insights around how "getting things done" may differ across contexts. For example, it may be suggested that successful policy enactment is more likely to require the involvement of local authorities, unions and professional associations in some contexts compared with others. From a methodological perspective, these insights are a feature of the comparative case study approach that we used in this project, and such insight could potentially be overlooked if analysis focused on any one single national case (Fitzsimons & Johnson, 2020).

Despite these differences, our analysis also allowed us to see some convergence. The common focus across all countries on ensuring that underprivileged learners are a priority for policy is a signifier of a cultural core that underpins the systems. The debate around pupil wellbeing shows some variation in the different countries, but also reflects a largely common discourse around a shared concern. Wellbeing concerns steer the cautious guidance to schools around using formative assessment. These areas of shared focus remind us of the substantive common policy values (Jeffery & Wincott, 2006) that link people across the UK, and which coalesce around things such as attitudes to the welfare state and the National Health Service. In this sense, the composition of the policy mix in the different countries had strong common features.

Complex interdependence

Our study methodology furthermore allows us to consider whether there are issues of interdependence between the countries' systems. The pandemic context makes explicit the common legal frameworks that link the countries' actions. For example, advice around international travel by the UK Government's Foreign and Commonwealth Office in July had an undoubted effect on planning for school trips for all UK countries. The early response to the pandemic demonstrates the degree of organisational coordination that existed across the countries at that time (although this coordination has diminished more recently). Each devolved administration has a Chief Medical Officer (CMO) who works with the CMO to the UK Government to provide coordinated advice to government departments in the four countries. At the beginning of the pandemic, ministers from the devolved administrations also attended meetings of new Ministerial Implementation Groups that were established to look at specific aspects of the coronavirus response.[5]

The degree of interdependence between the devolved systems is also brought into sharp focus by the events around exam grading that occurred in the late summer of 2020 (which fell outside the time frame of consideration for the analyses that we include in this paper). The fallout from the exam grading decisions initially made in Scotland had repercussions on all of the other countries, suggesting that interdependence works in

5 The Ministerial Implementation Groups (MIGs) ceased to operate by early June 2020, and it appears that ministers from the devolved administrations have not been part of the UK Government's cabinet committees that replaced the MIGs (Paun et al., 2020).

other ways than through direct policy collaboration. Interdependence in this sense may also account for some of the fast policy response observed in our analysis, and recalls Boin et al.'s (2018) observation that flexibility and improvisation is a characteristic of decision-making in times of crisis. It also reinforces the need to consider the interconnections that exist in public discourse and public opinion across the borders, that is, looking beyond the connections that exist at policymaking level. Additionally, the influence of powerful stakeholders across national borders is particularly noticeable in a context of complex interdependence. While our analysis focused on evidence of impact on policymaking processes and implemented policies, there is a clear indication that while stakeholder engagement plays a role in the formation of policy "upfront", stakeholder reactions to policy and proposals also feed into the policy cycle from below. In areas of policy decision-making that are still in the balance, such as arrangements for 2021 examinations, stakeholder voices may prove highly influential both within national borders but also in public debate and decision-making in the other countries.

Methodological reflections

Our study reinforces to us the importance of recognising some issues that need to be considered when making comparisons across the UK countries. There are a number of conditions that differ across the countries that can help to explain why there are variances in policy formation and content. For example, there were differences in the extent of physical school closure across the different countries which would be expected to influence the need for curriculum support. The countries are at different stages of their curriculum trajectories (e.g., ranging from being relatively well established in England, to a process of "disestablishing" in Wales) and so have a different relationship to the curriculum.

A criticism of the comparative case study method that we employ for the type of analysis that we have used here is that it involves a degree of trade-off: foregoing a deep analytic focus on one national "case" in order to gain insights from across a broad array of national situations. However, the method utilises a form of analysis that is found across a diverse set of research methods, from Randomised Controlled Trials (RCT) to Ethnography. Where RCTs manipulate variables and look for direct comparison of outcomes in relation to those specific manipulations, Ethnography involves analysts reflecting on the special insights that they gain from possessing both "insider" and "outsider" perspectives. We argue, in line with Raffe and colleagues (1999), that the comparative framework that supports cross-national case study analysis provides useful insights if the overlaps between the chosen cases are sufficiently strong—which appears to be the case across the countries of the UK.

References

Boin, A., Hart, P. T., & Kuipers, S. (2018). The Crisis Approach. In *Handbooks of Sociology and Social Research. Handbook of Disaster Research* (pp.23–38). Springer. https://doi.org/10.1007/978-3-319-63254-4_2

Broekema, W. (2016). Crisis-induced learning and issue politicization in the EU: the Braer, Sea Empress, Erika, and Prestige oil spill disasters. *Public Administration, 94*(2), 381–398. https://doi.org/10.1111/padm.12170

Bryman, A. (2016). *Social research methods (5th ed.).* Oxford University Press.

Cairney, P. (2013, September). *'Majoritarian Versus Consensus Policymaking' or a 'Common European Policy Style'? The Case of the UK and Scotland.* Presented at the Japanese Political Science Association Annual Conference, Hokkai Gakuen University, Sapporo.

Capano, G., Howlett, M., Jarvis, D. S. L., Ramesh, M., & Goyal, N. (2020). Mobilizing Policy (In)Capacity to Fight COVID-19: Understanding Variations in State Responses. *Policy and Society, 39*(3), 285–308. https://doi.org/10.1080/14494035.2020.1787628

Coventry City Council. (2020). *Recovery Curriculum.* https://www.coventry.gov.uk/info/62/coventry_send_support_service/3631/recovery_curriculum

Croxford, L., & Raffe, D. (2014). Social class, ethnicity and access to higher education in the four countries of the UK: 1996–2010. *International Journal of Lifelong Education, 33*(1), 77–95. https://doi.org/10.1080/02601370.2013.873214

Department for Education for England. (2020a). *Supporting vulnerable children and young people during the coronavirus (COVID-19) outbreak.* https://www.gov.uk/government/publications/coronavirus-covid-19-guidance-on-vulnerable-children-and-young-people

Department for Education for England. (2020b). *Get help with technology during coronavirus (COVID-19).* https://www.gov.uk/guidance/get-help-with-technology-for-remote-education-during-coronavirus-covid-19

Department for Education for England. (2020c). *Guidance for full opening: schools.* https://www.education-ni.gov.uk/sites/default/files/publications/education/circular%20curriculum%20planning%20202021.pdf

Department of Education for Northern Ireland. (2020a). *Curriculum Planning 2020/21.* https://www.education-ni.gov.uk/sites/default/files/publications/education/circular%20curriculum%20planning%20202021.pdf

Department of Education for Northern Ireland. (2020b). *Weir publishes guidance on new school day.* https://www.education-ni.gov.uk/news/weir-publishes-guidance-new-school-day-0

Donnelly, C., & Osborne, R. D. (2005). Devolution, Social Policy and Education: Some Observations from Northern Ireland. *Social Policy & Society, 4*(2), 147–156. https://www.cambridge.org/core/journals/social-policy-and-society/article/devolution-social-policy-and-education-some-observations-from-northern-ireland/D95526324FE69E82AB5148378DD460F2

Dyer, W. G., & Wilkins, A. L. (1991). Better Stories, Not Better Constructs, to Generate Better Theory: A Rejoinder to Eisenhardt. *The Academy of Management Review, 16*(3), 613–619. JSTOR. https://doi.org/10.2307/258920

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *The Academy of Management Review, 14*(4), 532–550. JSTOR. https://doi.org/10.2307/258557

Estyn. (2020). *What would inspectors do in the transition year 2020-2021?* https://www.estyn.gov.wales/faq/what-would-inspectors-do-transition-year-2020-2021

Fitzsimons, S., & Johnson, M. (2020). How Collaborative Project Development Theory Can Be Used to Provide Guidance for International Curriculum Partnerships. *International Dialogues on Education: Past and Present, 7*(2), 24–39.

Glasgow City Council (2020, June). *Education Services: Framework for Recovery, Resilience and Re-connection (RRR2020).*

Grødem, A. S., & Hippe, J. M. (2019). The expertise of politicians and their role in epistemic communities. *Policy and Politics, 47*(4), 561–577.

Heale, R., & Twycross, A. (2018). What is a case study? *Evidence-Based Nursing, 21*(1), 7–8. https://doi.org/10.1136/eb-2017-102845

Howlett, M. (2019). Procedural Policy Tools and the Temporal Dimensions of Policy Design: Resilience, Robustness and the Sequencing of Policy Mixes. *International Review of Public Policy, 1*(1), 27–45.

Howlett, M., & Tosun, J. (Eds.). (2019). *Policy Styles and Policy-Making: Exploring the Linkages* (1st ed.). Routledge.

Jeffery, C., & Wincott, D. (2006). Devolution in the United Kingdom: Statehood and Citizenship in Transition. *Publius, 36*(1), 3–18. JSTOR. https://www.jstor.org/stable/20184939

Keating, M. (2010). *The government of Scotland: Public policy making after devolution* (2nd ed.). Edinburgh University Press.

Knill, C., Schulze, K., & Tosun, J. (2012). Regulatory policy outputs and impacts: Exploring a complex relationship. *Regulation & Governance, 6*(4), 427–444. https://doi.org/10.1111/j.1748-5991.2012.01150.x

Laming, D. (2004). *Human Judgment: The Eye of the Beholder.* Thomson Learning.

Lidher, S., McIntosh, M., & Alexander, C. (2020). Our migration story: history, the national curriculum, and re-narrating the British nation. *Journal of Ethnic and Migration Studies,* 1–17. https://doi.org/10.1080/1369183X.2020.1812279

Machin, S., McNally, S., & Wyness, G. (2013). Educational attainment across the UK nations: performance, inequality and evidence. *Educational Research, 55*(2), 139–164. https://doi.org/10.1080/00131881.2013.801242

Masri, A. E., & Sabzalieva, E. (2020). Dealing with disruption, rethinking recovery: Policy responses to the COVID-19 pandemic in higher education. *Policy Design and Practice, 3*(3), 312–333. https://doi.org/10.1080/25741292.2020.1813359

Mintrom, M., & O'Connor, R. (2020). The importance of policy narrative: effective government responses to Covid-19. *Policy Design and Practice*, 205–227. https://doi.org/10.1080/25741292.2020.1813358

NAHT. (2020). *Update on school reopening plans.* https://www.naht.org.uk/members-home/naht-cymru/naht-cymru-news/update-on-school-reopening-plans/

NI Direct. (2020). *Coronavirus (COVID-19): advice on schools, colleges and universities.* https://www.nidirect.gov.uk/articles/coronavirus-covid-19-advice-schools-colleges-and-universities

O'Brien, C. (2020, September 16). School Covid-19 safety check results will not be shared with parents. *The Irish Times.* https://www.irishtimes.com/news/education/school-covid-19-safety-check-results-will-not-be-shared-with-parents-1.4355817

Ofqual. (2020). *Consultation Decisions: An additional GCSE, AS and A level exam series in autumn 2020.*

Paterson, L., & Ianelli, C. (2007). Social Class and Educational Attainment: A Comparative Study of England, Wales, and Scotland. *Sociology of Education, 80,* 330–358.

Paun, A., Rutter, J., & Nicholl, A. (2016). *Devolution as a Policy Laboratory: Evidence Sharing and Learning Between the UK's Four Governments.* Alliance for Useful Evidence. https://www.instituteforgovernment.org.uk/sites/default/files/publications/Alliance%20Policy%20Laboratory%20paper%20v3.pdf

Paun, A., Shuttleworth, K., Nice, A., & Sargeant, J. (2020). *Coronavirus and devolution.* Institute for Government. https://www.instituteforgovernment.org.uk/explainers/coronavirus-and-devolution

Professional Association for Childcare and Early Years. (2020, July 3). *NEWS: Ofsted inspections in England to fully resume from January 2021.* https://www.pacey.org.uk/news-and-views/news/archive/2020-news/july-2020/news-ofsted-inspections-in-england-to-fully-resum/

Public Health Wales. (2020). *Putting health and wellbeing at the core of the new Welsh education curriculum.* https://phw.nhs.wales/news/putting-health-and-wellbeing-at-the-core-of-the-new-welsh-education-curriculum/

Qualifications Wales. (2020). *Standardisation arrangements for summer 2020 in Wales.* https://www.qualificationswales.org/media/6353/standardisation-arrangements-for-summer-2020-in-wales.pdf

Raffe, D. (2005, March 15). *Learning from 'Home International' Comparisons: 14-19 Curriculum and Qualifications Reform in England, Scotland and Wales.* Presented at the Joint seminar of Education and Youth Transitions Project and Nuffield Review of 14-19 Education, London, UK. http://www.ces.ed.ac.uk/old_site/PDF%20Files/EYT_SP04_DR.pdf

Raffe, D., Brannen, K., Croxford, L., & Martin, C. (1999). Comparing England, Scotland, Wales and Northern Ireland: The case for "home internationals" in comparative research. *Comparative Education, 35*(1), 9–25. https://doi.org/10.1080/03050069928044

Richardson, J. J. (1982). *Policy styles in Western Europe.* Allen & Unwin.

Scottish Government. (2020a). *COVID-19 Education Recovery Group*. https://www.gov.scot/groups/covid-19-education-recovery-group/

Scottish Government. (2020b). *Coronavirus (COVID-19): strategic framework for reopening schools, early learning and childcare provision.* https://www.gov.scot/publications/excellent-equity-during-covid-19-pandemic-strategic-framework-reopening-schools-early-learning-childcare-provision-scotland/

Scottish Government. (2020c). *Coronavirus (COVID-19): Curriculum for Excellence in the Recovery Phase.* https://www.gov.scot/publications/coronavirus-covid-19-curriculum-for-excellence-in-the-recovery-phase/

Scottish Government. (2020d). *Coronavirus (COVID-19): statement by the Deputy First Minister on re-opening of schools.* https://www.gov.scot/publications/re-opening-schools/

Scottish Government. (2020e). *Coronavirus (COVID-19): support for continuity in learning.* https://www.gov.scot/publications/coronavirus-covid-19-support-for-continuity-in-learning/

Scottish Government. (2020f, September 2). *Coronavirus (COVID-19): guidance for school visits and trips.* https://www.gov.scot/publications/coronavirus-covid-19-guidance-for-school-visits-and-trips/

Scottish Qualifications Authority. (2020). *Post-Certification Review: Information for centres.* https://www.sqa.org.uk/sqa/files_ccc/post-certification-review-information-for-centres.pdf

Sibieta, L., & Cottell, J. (2020). *Education policy responses across the UK to the pandemic.* Education Policy Institute. https://epi.org.uk/publications-and-research/education-responses-uk-pandemic/

Simkins, T. (2014). Governance of Education in the United Kingdom: Convergence or Divergence? *Local Government Studies, 40*(6), 986–1002. https://doi.org/10.1080/03003930.2012.726143

TES. (2020, June 2). *School inspections in Scotland suspended next year.* https://www.tes.com/news/coronavirus-school-inspection-scotland-suspended-next-year

The Institute for Government. (2019). *Devolution at 20.* https://www.instituteforgovernment.org.uk/publication/devolution-at-20/overview

TUC. (2020). *Unions set out needs for safe reopening of schools in letter to Education Secretary*. https://www.tuc.org.uk/news/unions-set-out-needs-safe-reopening-schools-letter-education-secretary

UK Government. (2020a, March 12). *Government advises schools and colleges to cancel overseas trips.* https://www.gov.uk/government/news/government-advises-schools-and-colleges-to-cancel-overseas-trips

UK Government. (2020b, March 23). *Foreign Secretary advises all British travellers to return to the UK now.* https://www.gov.uk/government/news/foreign-secretary-advises-all-british-travellers-to-return-to-the-uk-now

UK Government. (2020c, July 3). *Coronavirus (COVID-19): travel corridors.* https://www.

gov.uk/guidance/coronavirus-covid-19-travel-corridors

Weale, S. (2020, June 9). Reopening schools: what is happening in England? *The Guardian*. https://www.theguardian.com/education/2020/jun/09/reopening-schools-what-is-happening-in-england

Welsh Government. (2020a). *Guidance on learning in schools and settings from the autumn term.* https://gov.wales/sites/default/files/publications/2020-07/guidance-learning-schools-settings-from-autumn-term.pdf

Welsh Government. (2020b). *CORONAVIRUS ACT 2020: Disapplication of Curriculum Requirements in Wales Notice 2020.* https://gov.wales/sites/default/files/publications/2020-06/disapplication-of-curriculum-requirements-in-wales-notice-2020.pdf

Welsh Government. (2020c, August 20). *Coronavirus: travel guidance for education settings and students (version 2).* https://gov.wales/coronavirus-travel-guidance-education-settings-and-students-html

Welsh Parliament Research Service (Senedd Research). (2020, April 8). *Coronavirus: schools and pupils.* https://seneddresearch.blog/2020/04/08/coronavirus-what-has-been-the-impact-on-schools-and-pupils/

Wilson, D. (2011). Comparative Analysis in Public Management. *Public Management Review, 13*(2), 293–308. https://doi.org/10.1080/14719037.2010.532967

Yin, R. K. (1981). The Case Study as a Serious Research Strategy. *Science Communication, 3*(1), 97–114. https://doi.org/10.1177/107554708100300106

Generation Covid and the impact of lockdown

Gill Elliott Research Division

Introduction

This article explores the different groups of school-aged students in England whose educational provision has been affected by the pandemic of 2020 and 2021. I look at who they are and the impact that lockdown educational arrangements might have upon subgroups within the cohort. I also consider the research implications arising from the need to evaluate these students' progress through the education system and beyond.

Our organisation is concerned with education in a wide sense. We have provided assessments for over 150 years, but over the past quarter-century this has expanded to include curricula and educational support. Our organisational mission is to unlock the power of education for every student. In line with that mission, we are working to evaluate and minimise the effects of the disruption from the pandemic upon the users of our products and services. This article seeks to provide additional insight into that process by providing a factual account of the school-aged year groups in England who are affected by the disruptions caused by COVID-19. It takes a step back from the very detailed analyses and debates that are ongoing in every educational sector about the details of arrangements and assessments in order to record, for posterity, an overview of the student groups who are affected.

Background

COVID-19 was declared a pandemic on 11 March 2020 (Cucinotta & Vanelli, 2020). From an educational perspective, in England, the first point at which students' lives began to change was with the onset of lockdown[1] on 20 March 2020, when schools were closed to all pupils except key workers' children and those deemed vulnerable[2] (Kim & Asbury, 2020) as a means of reducing the spread of the virus (Viner et al., 2020). At this point, learning moved online with immediate effect, in all circumstances where it possibly could. Primary school students, parents and carers carried out home-schooling, using activities provided by the school. Both caregivers and schools supplemented these activities with

1 Lockdown is the term used widely to refer to the closure of non-essential retail services and entertainment/leisure events and venues which was combined with instructions to the public to stay at home whenever possible. At various points in time, there were also restrictions on distance that might be travelled, activities which could be undertaken out of doors and fines for breaching these restrictions. At times when full national lockdown was not in place, a tier system was also used, which allowed different freedom of movement in geographical regions according to the proportion of the population infected.

2 The definition of vulnerable students includes those assessed as being in need under section 17 of the Children Act 1989, including children and young people who have a child in need plan, a child protection plan, or who are a looked-after child, or who have an education, health and care (EHC) plan, or who have been identified as otherwise vulnerable by educational providers or local authorities (including children's social care services). For further examples, see GOV.UK (2021a).

material provided via television or online sources. Secondary school-aged students were largely taught online by their school teachers, and studied independently. By 30 April 2020, the World Bank (The World Bank, 2020) reported that 674,868,288 students were in countries with fully closed schools and a further 93,602,500 in countries with partially closed schools. England fell under the latter category.

English schools take a long summer break from the middle of July to the beginning of September. In 2020, this coincided with the gradual easing of lockdown with the result that September 2020 saw most students returning to schools, albeit under arrangements which restricted unnecessary mixing, distanced pupils and mandated mask-wearing in some situations. Disruption still prevailed, with considerable numbers of students self-isolating[3] at any given time, and, in some cases, 50 per cent of lessons continuing to be delivered online in order to keep year group bubbles[4] separate.

The emergence of a new, more infectious variant of the coronavirus in late 2020 meant that the Government's aim to keep schools open carried too much risk. After a Christmas break that many students will have found disturbingly unfamiliar and stressful, a second lockdown shut schools on 5 January 2021, after one day back. Alongside the announcement of this was the notice that the examination session of summer 2021 was cancelled.[5]

Intention and limits of this article

I have restricted the scope of this article to England. The discussion points raised will certainly apply to students in other parts of the UK nations, and to those in other countries, but everyone's context is different and it is beyond the scope of this piece to include examples from other jurisdictions.

I have also limited my discussion to the effects of the events of March 2020 to July 2021. There seems good reason to do so. As this piece is being written, a vaccine is being rolled out and there is general optimism that disruptions from lockdowns or other government interventions will be minimal by the school year which begins in September 2021. So, although all current school year groups feature in the tables, I am assuming that the level of actual disruption in schools seen in the two academic years from September 2019

3 Self-isolation is the mandatory restriction of a person to their home if they have the illness or have been in close contact with someone who has it.

4 Bubbles refer to a defined cluster of people who are grouped together and who remain separate from others. In a domestic setting, a bubble could consist of more than one household, if certain criteria were met. In schools, this means groups of students who are separated (by space or over time) even when sharing the same facilities. For example, in a sixth form consisting of groups, each year group could be made into a separate bubble and would attend college premises on alternate weeks to minimise contact between the two groups. The principal rule which applies to bubbles is that if one person in the bubble contracts COVID-19, the whole bubble has to self-isolate.

5 For absolute clarity and accuracy, it is a point of record that the Government announcement (Johnson, 2021) stated that "it is not possible or fair for all exams to go ahead". The interpretation of the phrase depends upon whether "all exams" is taken to mean "any exams" or just "some exams". The media interpretation, shared widely and immediately, was that the intended meaning was no exams would go ahead.

to July 2021 will have abated by the beginning of the school year in September 2021. If history proves me wrong, which I sincerely hope it does not, the reader may be able to deduce the effects upon successive cohorts, depending upon where they have progressed in the system.

In this article, for simplicity, I use the terminology of "schools" to refer to places of education up to the age of 18, "assessment" to refer to forms of formal qualification or certification and "students" to refer to the learner population. Further explanation of the terms can be found in the endnote.[i]

Affected students in England by year group

Table 1 shows the school year groups in England,[6] all of which will be affected, to a greater or lesser extent, by the events of 2020 and 2021. The table shows their birth year and the year group they belonged to in March 2020, the latter expressed using the English education convention of "Y" to denote "year" followed by the number of the school year, which ranges from "-1" denoting the year prior to "R" (reception year, effectively year 0 in the school system) to 13 (the second and final year of sixth form for many students following a traditional academic route).

Table 1 also indicates particular features of concern relevant to that year group and whether key educational assessments were due within the year. For brevity in the discussions which follow, GCSE (General Certificate of Secondary Education), IGCSE (International General Certificate of Secondary Education) and vocational assessments taken at the end of Y11 are described as 16+ assessments; A Level, IB and vocational assessments taken at the end of Y13 are described as 18+ assessments—those being the typical age at which they are sat.

Table 1 also captures the broad extent of the non-compulsory curricular events that were likely to have been curtailed. The latter are important because many students will have suffered the loss of a great many peripheral activities that would normally have been valuable additions to their first or early-career curriculum vitae. Considering that many of the part-time and casual employment opportunities for the 15–18 age group are in the hospitality and retail industries, which have been especially hard-hit by lockdown, it becomes clear that Further and Higher Educational Institutions and employers will need to alter their expectations of what experiences they will see in applications for many years to come. Older students in this table will be under-experienced in many parts of their lives, not just their educational progress. There is also a possibility that some of the younger students will miss opportunities for learning and being assessed in skills such as cycling proficiency and swimming, which may result in an increase in accidents among this cohort in years to come.

What we see from Table 1 is that, in general, some year groups are likely to be more affected than others. Broadly speaking, it is those students who are approaching the end of compulsory education or those at transition points whose assessments and onward trajectory are most likely to be affected. However, it is also clear that students across

6 Compulsory education only. Excluding university students and pre-school provision earlier than the year immediately before compulsory primary attendance.

all year groups will face issues connected to particular areas of lost learning or lost school experience.

Table 1: Identifying the Covid generation in England.

Birth year 1 Sep–31 Aug	School year in March 2020	Particular concerns about this cohort	Educational content and assessments affected (content from National Curriculum)[7]	Educational assessments/ achievements affected—non-compulsory curriculum
2001/02	13	Final year of sixth form or Further Education College studies or learning within apprenticeships. University entry applications are often completed by this year group.	High-stakes 18+ assessments summer 2020.	• Duke of Edinburgh programme; • Music assessments—e.g., ABRSM, Trinity College and others; • Dance assessments; • Other sports, arts and academic awards and competitions—e.g., Maths Olympiad, public speaking, school plays, sporting fixtures; • Prefect-ships and school teams captaincy; • Some apprenticeship programmes disrupted—either the college-based component or the on-the-job training or both. Students furloughed, or prevented from travelling (Ventura, 2020).
2002/03	12	First year of sixth form studies or Further Education College studies or learning within apprenticeships.	High-stakes 18+ assessments summer 2021.	
2003/04	11	Final year of secondary school-based studies. Sixth form entry. This year group will have had 16+ assessments replaced by Centre Assessed Grades (CAGs) and 18+ assessment courses disrupted for at least 2 out of 6 terms.	High-stakes 16+ assessments summer 2020. High-stakes 18+ assessments summer 2022.	
2004/05	10	First year of GCSE courses and many vocational pathways.	High-stakes 16+ assessments summer 2021. High-stakes 18+ assessments summer 2023.	

7 GOV.UK (2021b).

Birth year 1 Sep–31 Aug	School year in March 2020	Particular concerns about this cohort	Educational content and assessments affected (content from National Curriculum)[7]	Educational assessments/ achievements affected—non-compulsory curriculum
2005/06	9	Transition choices year from general to specialised curriculum.	Key Stage 3 curriculum. High-stakes 16+ assessments summer 2022.	• Music assessments—e.g., ABRSM, Trinity College and others; • Privately funded, school-based music tuition opportunities lost; • British Gymnastics Award Schemes.
2006/07	8		Key Stage 3 curriculum.	
2007/08	7		Key Stage 3 curriculum. Common Entrance Exam (Independent School Entry) in 2020.	
2008/09	6	Transition to secondary school.	Key Stage 2 curriculum. Key Stage 2 national assessments in 2020. Common Entrance Exam (Independent School Entry) in 2021.	• Department for Transport Bikeability scheme; • Swim England swimming challenge awards; • End of primary celebrations and events; • Privately funded, school-based music tuition opportunities lost; • Music assessments—e.g., ABRSM, Trinity College and others.
2009/10	5		Key Stage 2 curriculum. Key Stage 2 national assessments in 2021.	• Music assessments—e.g., ABRSM, Trinity College and others; • Privately funded, school-based music tuition opportunities lost.
2010/11	4		Key Stage 2 curriculum. Key Stage 2 national assessments in 2022.	
2011/12	3		Key Stage 2 curriculum.	

Birth year 1 Sep–31 Aug	School year in March 2020	Particular concerns about this cohort	Educational content and assessments affected (content from National Curriculum)[7]	Educational assessments/ achievements affected—non-compulsory curriculum
2012/13	2		Key Stage 1 curriculum. National tests and teacher assessments in Maths, English and Science.	
2013/14	1		Key Stage 1 curriculum. Phonics screening check.	
2014/15	R	Just settled into Reception as lockdown in 2020 began. Restarted in September 2020 only to be locked down again in January 2021.	Early years curriculum. Teacher assessments.	• Introduction to school life very severely disrupted.
2015/16	-1 Pre-school	Transition to primary school from home or pre-school occurring just after lockdown, with limited preparatory events. Locked down again in January 2021.	Early years curriculum.	• Introduction to school life very severely disrupted.

Learning loss has not been specifically captured by this table—aside from Y13, the extent of learning loss (in terms of hours attending school) is assumed to be broadly the same across all year groups, even if it differs between educational settings. Year 13 lost a few weeks of their teaching/revision time, but had nearly completed face-to-face teaching and learning at the point of lockdown. In other year groups, the content of the learning loss will vary between year group and setting and is beyond the scope of this discussion.

Learning loss is also only half of the story. While there is a rightful focus on lost learning and how it affects the students approaching high-stakes assessments, the impact of the pandemic has been keenly felt in the loss of the school environment. Schools are not just places where students are taught. They are a cultural and social hub, where students experience consistent attitudes to learning and are given opportunities that do not

necessarily exist within the home environment.

While Table 1 shows the core information about who is part of the Covid generation of students and the key educational and assessment milestones that have been very specifically disrupted, it is important to delve deeper to look at what impact might be felt and how this might differ for different subgroups of students.

What complications systematically affect the impact of COVID-19 disruption for different groups of students?

In Table 1, we saw how the core year groups of students within compulsory education in England have been disrupted by COVID-19. However, we cannot assume that this disruption occurred in the same way for all students.

The English education and assessment system is arguably built upon a principle of meritocracy. Put very simply, this means that progress and achievement is based upon individual performance, talent and effort rather than wealth or social class. In reality, other sociological factors do play a part and there has been much argument about whether equality of educational opportunity exists (Heaton & Lawson, 1996, pp.146–153). Nevertheless, key to the system, as it normally functions, is a principle of equality of opportunity regardless of factors such as gender, class, background, race, religion or economic circumstances. However, viruses are no respecters of educational principles and closing schools in order to reduce the spread of the virus has had effects which exacerbate the impact of disadvantage for many students:

(i) The underlying principle that all students have the same equality of opportunity for learning does not hold when you look at the cohort as a whole. Under normal circumstances, in England, we can assume that every student (with the exception of those noted in (ii) for whom mitigations are applied) has a similar basic availability of qualified teacher-led learning hours. The amount of learning loss during the pandemic cannot be assumed to be equally distributed across all students because it depends upon an interaction between each individual student's home circumstances and their school's response.

(ii) Many of the methods by which equality of opportunity is traditionally managed are dependent upon attendance in person at school; for example, free school meals to compensate students who would otherwise struggle to learn when physically poorly nourished. While schools have remained open for vulnerable students, and communities have, in some cases, provided alternate provision, these arrangements cannot be presumed to be applied equally.

(iii) New factors affecting equality of opportunity have emerged and it is unlikely that these can be fully documented or quantified in time for the alternate arrangements for assessments in summer 2021, even though Y10 and Y12 students taking two-year courses such as GCSEs, IGCSEs and A Levels will have been struggling with them for five out of the six terms[8] of their course of learning. Y9 and Y11 students taking assessments in 2022 will also have been affected for at least two of the six terms of

8 The English school system is divided into three terms per academic year.

their courses (possibly more, because at the time of writing the arrangements for the summer term of 2021 are not yet known). These new factors include, but are not limited to:

- difficulties with technology including, but not limited to, insufficient internet bandwidth to support full engagement with some or all forms of online learning organised by the school and differences experienced at school level with the organisation or provision of an effective online experience;
- non-conducive learning environments outside the student's control;
- less motivation to learn online (which the student cannot control for themselves if they do not understand why it is happening);
- more distractions, anxieties and mental health issues;
- inability to complete the full course of learning; and
- unwillingness to expose their home environment to scrutiny via video camera, thus less face-to-face contact.

Some of these issues (the second to penultimate points above) will also apply to vulnerable students attending school settings, who are being asked to attend lessons online from the school setting, that is, if all attending students are supervised together in a classroom and each follows their individual lesson programme online.

Table 2 presents, in a structured way, some of the different issues which have arisen. However, one further complication when considering the nature of disadvantage is that some students have greater opportunities than others by virtue of their particular school or home environment.

England has a broadly two-track education system with selective or fee-paying schools as well as state schools. It would be naïve to argue that fee-paying or selective schools automatically advantage their students (in normal times) by comparison with state schools, because, at individual level, students in either environment can make more or less use of the resources they are offered. Nevertheless, a quick survey of the claims made on a few fee-paying schools' websites suggests a number of areas where their students are claimed to benefit from enhanced opportunity:

- Enriched educational experience, through extracurricular initiatives, programmes outside the curriculum and the opportunity to set their own curriculum and tailor learning beyond the boundaries imposed in the state sector;
- Reduced class size—more one-to-one time;
- Safe learning environment—higher discipline; and
- Excellent resources (technical, sports, facilities).

Advantages due to enhanced resources, as described above, have direct relevance to the very specific issues which have arisen around coping with learning during lockdown. Online learning works best with efficient technology, smaller groups and individual support for students who struggle with it. Therefore, in any attempt to systematically

consider how lockdown has affected different groups of students, some of the factors which characterise (but are not inclusive to) fee-paying or selective schools must be considered. However, these factors should not be regarded as entirely indicative of particular school types—in the context of "lockdown learning", a parent who is an experienced IT user and is working from home might be just as valuable and effective a resource to a student as enrolment at a school which has enhanced IT solutions available. Therefore, in Table 2, three categories of students are shown which are not based on school type: Group 1—neither advantaged nor disadvantaged normally, but affected by lockdown; Group 2—advantaged in some way that mitigates the impact of lockdown and Group 3—disadvantaged in the normal system and with additional disruption from lockdown.

The categories in Table 2 are not mutually exclusive for individuals. For example, it is perfectly possible for a student to fall into the category of being advantaged by virtue of enhanced personal financial resources and also fall into the category of being disadvantaged by not coping well with traditional assessment models.

A key factor of importance is whether the individual student has control of the advantage or disadvantage being described. A student who has spent most of lockdown playing Call of Duty on their Xbox to the detriment of their studies has still had the opportunity to engage with online learning, whereas one who has been repeatedly unable to access sufficient bandwidth to participate in lessons has not.

Table 2: Effects of lockdown on individual subpopulations of students.

	Nature of the disadvantage or advantage	How is this reflected?	What differences will the pandemic, and particularly the locked-down environments from March–May and in November 2020 have made to their educational experience?
Group 1—neither advantaged nor disadvantaged normally, but affected by lockdown.	No particular advantage or disadvantage. All have broadly the same opportunities to learn, similar pathway choices, and a similar set of assessment experiences.	Individual experiences will vary.	With the possible exception of Y13 (births 2001/02), all students have experienced severe learning loss during their compulsory school education. Y13 had almost finished their taught courses in March 2020, but will have missed school-based face-to-face revision and assessment preparation sessions. Many Y13 students taking practical assessments requiring school-based facilities may have been unable to fully complete them. Online learning approaches will have been of mixed success across this group, largely due to the need to overcome practical issues and to acclimatise teachers and students with the differences in approach.[9]

9 Some of these students may have been identified as vulnerable—the Government definition (GOV.UK, 2021a) includes as an example "those who may have difficulty engaging with remote education at home (for example due to a lack of devices or quiet space to study)"—and offered a part- or full-time place in school.

	Nature of the disadvantage or advantage	How is this reflected?	What differences will the pandemic, and particularly the locked-down environments from March–May and in November 2020 have made to their educational experience?
Group 2—advantaged in some way that mitigates the impact of lockdown.	Enhanced personal financial or non-financial resources, either of which leads to an enriched cultural and educational experience, either in an educational setting or outside of it.	Greater opportunity to attend a fee-paying school environment. Greater access to equipment and software. Provision of additional private tuition. Greater support for learning. Members of household have skills which support online learning.	As Group 1, although online learning approach may have benefited from greater support or investment. Existing differences in achievement and attainment between this group and Group 1 will be increased.
	Attending an educational establishment with greater financial resources than the norm.	Enriched educational experience. Additional technology.	
	Attending an educational establishment with greater non- financial resources than the norm.	More tailored learning environment.	
	Advantaged by lockdown.	Some students will, for a number of reasons, thrive in the alternative learning environment that lockdown creates.	These students will achieve greater educational progress than they might have done under normal circumstances and will attain better results in assessments.
Group 3—disadvantaged in the normal system and with additional disruption from lockdown.	Students who have underperformed at an earlier stage of their education and been unable to catch up.	Tailored learning approaches in schools under normal circumstances.	There may be more variety in the provision of support for these students and it is difficult to predict which students will become more disadvantaged and which might regain some progress.

	Nature of the disadvantage or advantage	How is this reflected?	What differences will the pandemic, and particularly the locked-down environments from March–May and in November 2020 have made to their educational experience?
	Students who do not cope well with traditional learning or assessment models.	Individual support in schools under normal circumstances.	Students who do not cope well with traditional teaching methods in schools are probably unlikely to find online teaching easier. Students who struggle with assessments are likely to suffer heightened anxiety around preparing for 2021 assessments in the atmosphere of "will they/won't they happen". There may be fewer resources available for support in schools.
	Students with underlying mental health issues.	Individual support in schools under normal circumstances.	Existing support for mental health issues may have been affected by lockdown. Existing mental health issues may be worsened or changed by the changes due to the ongoing situation and the loss of control over daily schedules. Many students will be experiencing mental health issues for the first time as a direct consequence of lockdown.
	Students who are not equipped with sufficient language skills (e.g., economic migrant, asylum seeker and refugee students).	Individual support in schools under normal circumstances.	Fewer resources available for support in schools.
	Students who require accommodations to be made in order to access education (e.g., Special Educational Needs and Disabilities).[10]	Individual learning programmes. Assessments with accessibility provided.	Increased stress, worry and other mental health issues in students and their families. Amplification of the effects of loss of routines, support networks and structures and of specialist input by comparison with other groups (Asbury et al., 2020).
	Students whose time and/or effort commitment is curtailed due to circumstances beyond their control (e.g., young carers).	Support from schools under normal circumstances.	Hard to predict as a number of underlying factors could alter. Lockdown might provide greater time, with less need to travel. Existing support services might be disrupted leading to increased distraction and worry or decreased time.

10 Many of these students will have been considered vulnerable and offered a full- or part-time place at school. Not all will have taken up a place where offered.

	Nature of the disadvantage or advantage	How is this reflected?	What differences will the pandemic, and particularly the locked-down environments from March–May and in November 2020 have made to their educational experience?
	Students from an educationally, culturally or socially disadvantaged background.	School settings provide stability, equipment and strong adult support under normal circumstances.	Likely to fall behind educationally. May not be able to participate in lessons, unless equipment provided. Even where equipment can be provided, may be unable to use it, maintain it or keep it safe.
	Students who have experienced unusual disruption in their lives at the time of assessment.	Special Consideration.	There will be many more of these students during a pandemic.
	Students who live in poverty.	School settings provide resources such as food and equipment under normal circumstances.	Some support provided by schools and wider community provision, but the effects will be larger than usual.

The need for research to support the Covid generation

Fulfilling the educational obligation to the Covid generation of students will include tracking their progress for years to come in order to ensure that they are as little disadvantaged as possible. It will also be important to ensure that the measures taken to protect them do not disadvantage non-Covid generations.

It is clear that there will be an ongoing and extensive educational research need, both:

i. to monitor and evaluate the effects of the educational disruption experienced by the year group cohorts from Table 1 as a whole; and
ii. to provide evidence which will contribute to the evaluation and mitigation of disproportionate effects or biases felt by subgroups such as those described in Table 2.

Within the assessment research community, there will also be additional concerns about managing high-stakes assessments. There are ongoing practical dilemmas about how best to provide an effective and fair assessment solution to successive year groups, each of which has significantly different experiences in terms of content and learning loss. There is a need to manage the expectations of those who use the results of high-stakes assessments for selection purposes and assist with their interpretation of results from different years. Finally, there are numerous technical difficulties to be explored and overcome. For example, understanding, tracking and explaining any differences in standard of high-stakes assessments which have occurred as a consequence of the practical solutions implemented for the year groups concerned will be critical. Planning ahead for cohorts who have missed earlier national assessments (e.g., Key Stage 2 testing)

will also be needed, if those results are used at later stages of assessment provision.

Figure 1 illustrates some of these cohorts (year groups) on a timeline.

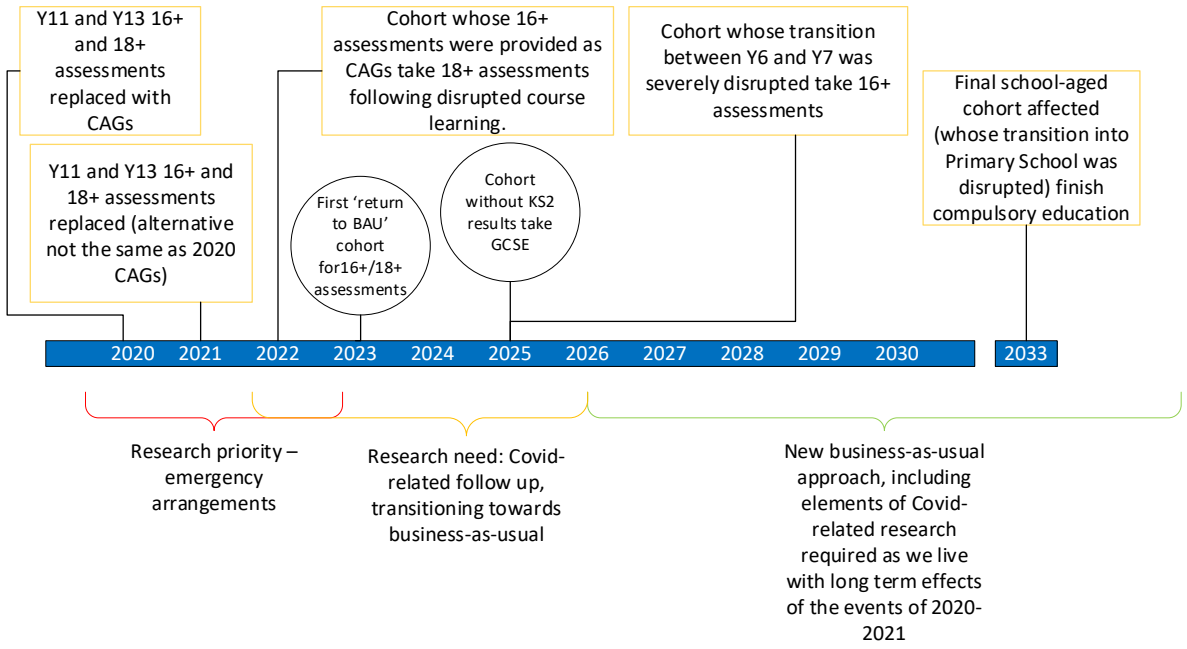


Figure 1: Timeline of research needs for different age cohorts.

Underneath the timeline is a suggestion about how the pandemic might impact upon assessment research specifically. The timescale is based upon the presumption that ongoing educational disruption will have ended by September 2021 (i.e., that students will experience normal school-based education from September 2021) and "business as usual" refers to research which is not concerned with pandemic-related arrangements. Ultimately, the two will blend together and the research response will move from an emergency "just in time" or reactive approach to a long-term business-as-usual strategy which incorporates ongoing elements of pandemic-related research.

Our particular interest is in those areas of research which focus upon high-stakes assessments—many other research bodies will also be tracking other aspects of lifelong progress and wellbeing. Many educational agencies will be undertaking work to support the Covid generation. There is a danger of duplication of effort, with different agencies only becoming aware of each other's work upon publication. There is also a risk that the sheer volume of research output will overrun the resources available to policymakers and compromise effective engagement with all of the potentially useful material. Meta-analyses and review documents will be important tools to avoid either of these problems and some form of centrally organised collaboration and registration of research may be desirable, following principles such as those suggested by The Royal Society (2019).

Summary

In this article, I have sought to explore who the Covid generation in English schools are, and to look at some of the different ways they may be disadvantaged. I have also acknowledged that some students may have benefited, either in terms of relative comparison with peers (i.e., not as badly disadvantaged) or in absolute terms (e.g., if lockdown learning positively suited their personality or situation).

I think it is hugely important to consider, at this stage, how research evidence will play a part in shaping the future of this cohort of students. While the full impact of lockdown on their education may not emerge until today's students are a long way into their careers, there is much we can do to support them throughout their journey.

[i]A note on terminology. Certain terms within this article are being used consistently, in order to improve the reader experience. While it might be more precise to use variations of terminology when describing certain situations, the nuances of these variations are often understood only by those whose expertise is in a very niche field. In the interests of reaching a wider audience I am therefore applying the following definitions:

- Students—refers to up to 18 year olds in any part of the system who are engaged in the process of learning in any setting (including within an apprenticeship programme in a workplace).
- School—a place of learning or a central place of assessment (more specifically known as a centre). In this article, "school" might mean a workplace context or a centre which makes entries to assessments.
- Assessment—all forms of assessment which lead to certification of a programme of learning or experience through some form of structured and controlled system to ensure that only the student's own performance is recognised. This includes examinations and tests which are, traditionally, invigilated sessions. It also includes vocational assessments, apprenticeship certification and other forms of deriving qualification grades, such as the Centre Assessed Grades used in 2020. The term "examination" is still used where it is part of a known phrase, such as "examination session".

References

Asbury, K., Fox, L., Deniz, E., Code, A., & Toseeb, U. (2020). How is COVID-19 affecting the mental health of children with special educational needs and disabilities and their families? *Journal of Autism and Developmental Disorders.* https://doi.org/10.1007/s10803-020-04577-2

Cucinotta, D., & Vanelli, M. (2020). WHO Declares COVID-19 a Pandemic. *Acta Biomed, 91*(1), 157–160. https://pubmed.ncbi.nlm.nih.gov/32191675/

GOV.UK. (2021a). *Children of critical workers and vulnerable children who can access schools or educational settings* [Guidance]. Updated 9 March 2021. https://www.gov.uk/government/publications/coronavirus-covid-19-maintaining-educational-provision/guidance-for-schools-colleges-and-local-authorities-on-maintaining-educational-provision

GOV.UK. (2021b). *The National Curriculum*. https://www.gov.uk/national-curriculum

Heaton T., & Lawson T. (1996). *Education and Training.* Palgrave. https://doi.org/10.1007/978-1-349-13883-8_7

Johnson, B. (2021, January 4). *Prime Minister's address to the nation.* GOV.UK. https://www.gov.uk/government/speeches/prime-ministers-address-to-the-nation-4-january-2021

Kim, L., & Asbury, K. (2020). 'Like a rug had been pulled from under you': The impact of COVID-19 on teachers in England during the first six weeks of the UK lockdown. *British Journal of Educational Psychology, 90*, 1062–1083. https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bjep.12381

The Royal Society. (2019). *Harnessing educational research: summary.* DES5757. https://royalsociety.org/-/media/policy/projects/rs-ba-educational-research/educational-research-summary.pdf

The World Bank. (2020, April 30). *World Bank Education and COVID-19.* https://www.worldbank.org/en/data/interactive/2020/03/24/world-bank-education-and-covid-19

Ventura, G. (2020). The future of apprenticeships is in the balance. LSE Business Review. http://eprints.lse.ac.uk/104776/1/businessreview_2020_05_02_the_future_of_apprenticeships_is_in_the.pdf

Viner, R. M., Russell, S. J., Croker, H., Packer, J., Ward, J., Stansfield, C., Mytton, O., Bonell, C., & Booy, R. (2020). School closure and management practices during coronavirus outbreaks including COVID-19: a rapid systematic review. *The Lancet Child & Adolescent Health, 4,* 397–404. http://dx.doi.org/10.1016/s2352-4642(20)30095-x

Disruption to school examinations in our past

Gillian Cooke Cambridge Assessment Archives and Heritage and **Gill Elliott** Research Division

In the 162-year history of Cambridge Assessment, the examination cancellations of 2020 were unprecedented. However, disruption of some kind to school examinations is not a new phenomenon, and many of the challenges brought by COVID-19 echo those we have faced in the past. By exploring and analysing historical events, as we have done on previous occasions (Cooke, 2020; Elliott, 2011; Newton, 2011) we can review lessons learned and shed light on perspectives to current issues.

Here, through a study of Cambridge Assessment Archives, we consider five events in our history which have affected our organisation. The two World Wars clearly loom particularly large, but our operations were also affected by partition in India in 1947 and the strikes in the United Kingdom (UK) of the 1970s. The fifth has possibly the most resonance with events of 2020—the flu pandemic of 1918.

Cambridge Assessment was set up as the "Local Examinations Syndicate" (henceforth Cambridge Syndicate), an outreach department of the University of Cambridge. At a time when children's education was neither compulsory nor widespread, the introduction of the Cambridge "Local Examinations" represented a specific move in England to meet demands from aspiring professionals seeking a benchmark for end-of-school attainment levels. The exams were managed by an executive committee of Syndicates, and they quickly expanded to include candidates from overseas. During the 20th century, these modest end-of-school assessments grew to become a mainstay of society, with a global reach into all areas of educational assessment. Further details are available in Raban (2008).

By 1914, end-of-school examinations had become more widespread and the Cambridge Syndicate was one of several UK exam boards. Despite its name, the "Cambridge Locals" were offered to candidates at centres in the UK and overseas. The centres were the locations (usually schools) which were approved to hold the examinations, and those held overseas were beginning to attract candidates other than those with a UK connection.

We shall begin by providing short descriptions of these events in relation to our organisation and outline the resources we have drawn on from the corporate archives. In the second part of this article, we look thematically across the events and explore the decisions and actions taken. Finally, we conclude by examining whether we can apply any lessons from these previous events to the current situation.

It should be borne in mind that we have resisted speculating too far beyond the information presented. Realistically, what was noted in the historical documents cited cannot be the full picture of any event; there will have been other actions and viewpoints either unrecorded or recorded elsewhere. In the same way, we could use multiple additional historical sources to attempt to contextualise each nugget of information, but this is clearly beyond the scope of this piece. We have intentionally targeted our inspection of documents to those concerning times of significant disruption and we intend that, by doing so, we will provide a certain amount of historical perspective to the current

situation from the viewpoint of one organisation.

The events

First World War

By 1914, the Cambridge Syndicate was an established organisation with annual candidate entries of around 24,000 and examinations were held at around 350 "local" centres worldwide. The main examinations session was in December and the most immediate impact of the war was the loss of staff, as many rushed to join the armed forces in autumn 1914, some of whom did not return. Other disruptions followed swiftly when bombardments from the sea at West Hartlepool, Scarborough and Whitby interrupted the examinations that December and, the following year, examination documentation bound for Jamaica was lost when the aptly named steamer *The Candidate* was torpedoed. The Annual Reports record long delays in the awarding process but, despite earlier events, the only direct loss of candidates' work was a batch of Higher Examination scripts which were on a ship lost off the Indian coast in June 1917.

Flu pandemic

The flu pandemic was a global catastrophe whose impact is often overshadowed by the First World War. Between March 1918 and March 1920, it claimed more than 50 million lives worldwide—over 200,000 in the UK, and most of them young. There was no governmental or co-ordinated national response in Britain and no National Health Service to treat patients, many of whom died just hours after becoming ill. Coming as it did, at the end of the First World War, the pandemic added to mounting exhaustion and financial distress, and the Cambridge Syndicate minutes reflect this, recording a heated discussion about the management of two centres which had been forced to close. In the end, four centres closed completely and overall candidate numbers dropped by 10 per cent.

Second World War

The impact of the Second World War on the lives of civilians was unprecedented and this is the most documented of all the disruptions faced by the organisation. War measures were introduced straightaway from 1940: the July overseas examinations were discontinued, and the "Junior Exams" withdrawn from the UK. Significantly, the Cambridge Syndicate also moved swiftly to continue its operations with reduced staffing. The legacy of the Cambridge Syndicate from this period comes from two areas, although the second of these only came to light after the war had ended. At home, "Special Circumstances" forms record examinations held during civilian bombing raids, while overseas, specially drafted examination regulations tell of unscheduled examinations held during hostilities for prisoners of war in internment camps.

Partition of India

Following the Indian Independence Act of 1947, the British Raj governance structure dissolved, and the former British India was separated into the two dominion states of India and Pakistan. The Cambridge Syndicate served candidates within and across the

two states. Partition was accompanied by a huge amount of local violence and the displacement of students along religious lines, so the Cambridge Syndicate had to make rapid changes to practical arrangements such as the despatch, transport and marking of scripts, and adjust the governance structure of the examinations to take account of the changing political landscape. Candidates who had been displaced were traced, new centres established and new Local Secretaries (responsible for groups of centres) were appointed where needed.

Strikes of the 1970s

By 1971, the total number of entries to the Cambridge Syndicate's examinations was over 430,000, over half of whom were candidates in East Africa, India, Malaysia and the Caribbean who mostly took examinations in November. That year, there was a national postal strike in the UK for seven weeks from January to March and the awarding process of the examinations under these conditions presented a major challenge. From 1972 to 1974, waves of power cuts affected the UK and inflation became a dominating concern. Devaluation of the pound in 1973 caused costs to spiral and the organisation battled to keep afloat, suffering six years of deficit in one decade.

Impacts of the events and historical context

In the following section, we look across the events and explore the decisions and actions taken, under five separate themes:

1. Organisational adaptability: Business as usual versus emergency measures.
2. Financial impacts: Costs, cost-cutting and unforeseen expenses.
3. Principles: The quest to uphold standards.
4. Support for candidates and staff.
5. Business practice: Short-term measures, aftermath and long-term changes.

The need to carry on as before was an overriding philosophy which had particular resonance during wartime. To bow to disruption caused by an enemy of the state would have demonstrated weakness so, as a result, a stoic attitude to all walks of life, including the running of school examinations, prevailed and became normalised throughout much of this period. By the standards of today, the confrontations and pursuit of examinations among the death and horror of the pandemic and Indian partition appear callous, but the attitudes of the time were set by war. Support for candidates, staff and examiners at these times appears to be similarly thin. By comparison, the strike and utility disruptions of the 1970s had none of the human tragedy but nevertheless elicited a sympathetic response from the Cambridge Syndicate, indicative of changing times.

The need for both to remain solvent and to uphold examination standards are more recognisable to readers today. Quite simply, without reference to standards, the examinations would have lost value and it is noteworthy that the "lesser" disruption of the 1970s includes the most reference to costs.

Although the examinations have survived since 1858, there have been many changes since then and it would be naïve to attribute expansion overseas and increasing nationalisation

of UK examinations to tumultuous events in history. The fact is, however, that whatever the impact of these events, they did not stall progress and development, and may even have acted as a stimulant.

1. Panic actions and business as usual

In spring 1917, the Cambridge Syndicate voted to accept a single German candidate entry for the Certificate of Proficiency Exam and "to carry on as before". This "business as usual" approach was demonstrated a few weeks later when a conference of Head Masters discussed curriculum changes, including the re-introduction of Domestic Science to the curriculum, "useful for girls who wished to be teachers", and whether novels were "good enough as subjects for examination" for boys.

War measures were, nevertheless, introduced; the Cambridge Syndicate appealed to retain staff in 1916 and began to waive late entry fees and consider the practicalities of making carbon copies of candidates' scripts in 1917. Reflecting the stoicism that was popular at the time, the Annual Reports indicate little disruption—the 1917 Annual Report declared that "during the War the number of candidates entered for the Local Examinations has been well maintained". Less so the following year, when exams had to be abandoned in some centres in Asia and Africa. Delays were common and expected; most notable were the December 1915 scripts from Lagos and the Gold Coast which took 22 months to reach Cambridge, via the USA!

Before the Armistice in 1918, the pandemic had begun, but it is not referenced in the Cambridge Syndicate minutes until it peaked in October, where the second of the two long references, entitled "Influenza over again" suggests it was a topic exhausted in conversation. The initial entry is emotive: "Whole centres shut down. Return fees en bloc. This is a precedent". The minute secretary noted "sharply contrasted views" and, below the second entry, took the unusual step of scribbling "Temper of meeting most unsatisfactory". This was clearly not business as usual, yet the December exams session did go ahead with 90 per cent of candidates in attendance.

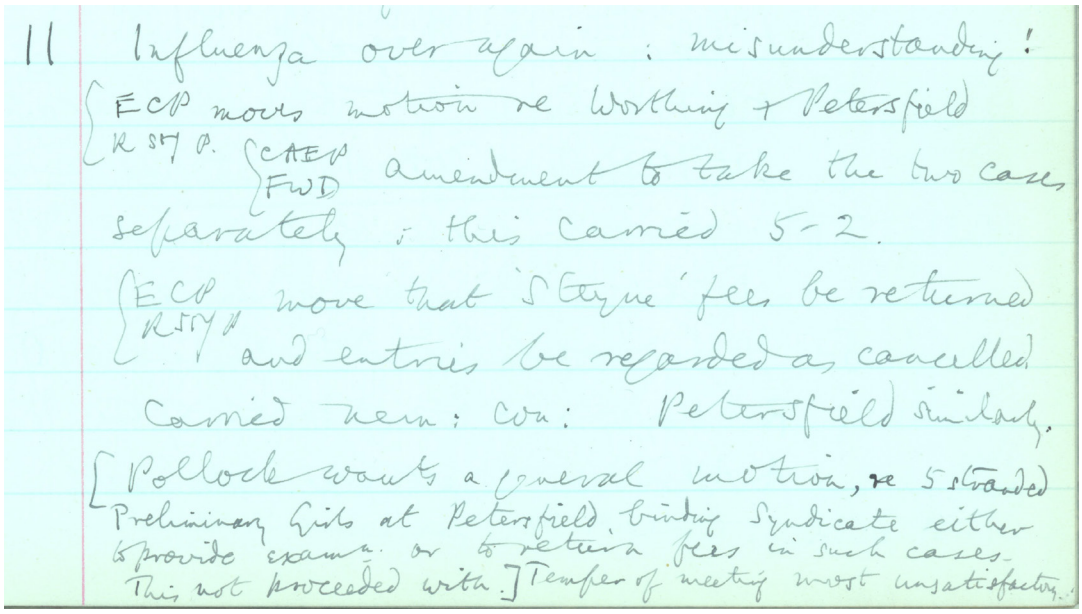

11 Influenza over again : misunderstanding!
ECP moves motion re Worthing + Petersfield
R STY P. CAEP FWD amendment to take the two cases
separately : this carried 5-2.
ECP RSTY move that 'Statute' fees be returned
and entries be regarded as cancelled
Carried nem: con: Petersfield similarly.
Pollock wants a general motion, re 5 stranded
Preliminary Girls at Petersfield, binding Syndicate either
to provide exams. or to return fees in such cases.
This not proceeded with.] Temper of meeting most unsatisfactory.

Figure 1: C/CB 1/2 Syndicate minutes 30 October 1918.

The early months of the Second World War are often referred to as the Phony War, characterised by uncertainty and a lack of military action. The abundance of documentation for this period includes a typescript memo entitled "Possible Emergency Arrangements" which outlines arrangements for evacuated schools in the event of an "international crisis". The Cambridge Syndicate sought to reassure centres and, the day following the outbreak of war, it wrote of its intention "to carry on … work in order to cause the minimum of disturbance to schools whether at home or overseas". With an acknowledgement to a previous era, the notice concludes with: "the situation will show whether it would be advisable to make duplicates of the December scripts before despatching them, as is believed to have been done in 1914–18". Over the following two years the Cambridge Syndicate discussed special regulations, special fees, emergency considerations and the appointment of fire watchers.

Centres affected directly by hostilities were encouraged to submit Special Circumstances reports to highlight candidates for special consideration and some of these still survive; they show that candidates whose homes were demolished, whose grandparents were killed, and who were fire-fighting all night, turned up to take exams the following morning and "showed a marked reluctance to single themselves out as undergoing special hardships" (Wintringham Secondary School, Grimsby). The examinations carried on as directed by the timetable; indeed, candidates at St James' Central School in Burnt Oak, Middlesex took their 1944 exams in an air raid shelter, with only four candidates singled out for special consideration, while candidates at Watford Central school "disregarded siren warnings" during exams and took shelter only when "danger imminent" signals were received.

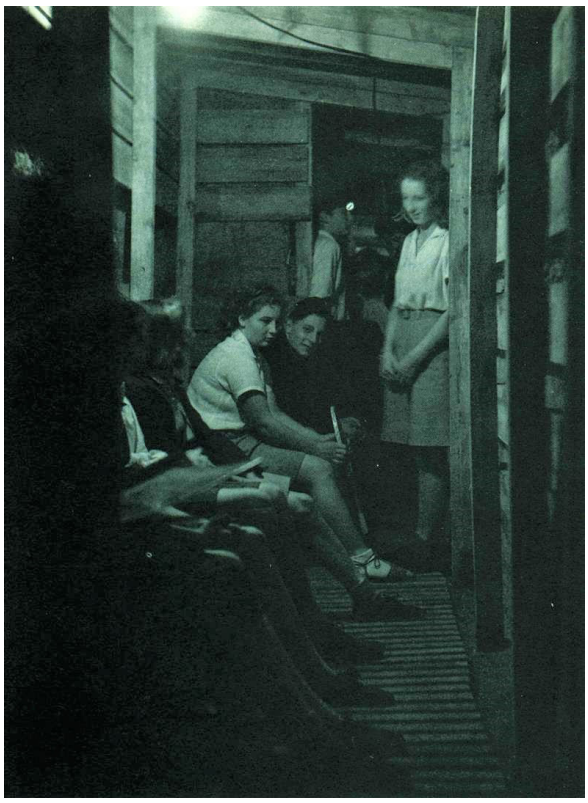


Figure 2: Bedales School air-raid shelter 1940. (Reproduced with kind permission of Bedales Archive.)

As with the previous war, the principal difficulties were transport delays, highlighted in the Annual Reports, which in 1942 record several overseas centres for which candidate scripts had not been received. 1946 marks a return to completed Annual Reports with a full breakdown of entries; it also includes a notification of the withdrawal of special allowances relating to war difficulties.

In similar vein, partition in India brought problems of transporting scripts, including frenzied efforts to secure alternative marking venues and the associated movement of the scripts to the new location. Difficulties were also noted about providing copies of texts, not least in tracking down the whereabouts of displaced publishers.

The Cambridge Syndicate minutes of January 1971 report that the impact of the strike in the UK was felt by over 4,000 examiners involved in post-examination work. Regional depots were set up in the UK, including one at the University of London Senate House, for examiners to deposit scripts for collection and transportation to Cambridge, and staff worked exceptional hours of overtime. The Cambridge Syndicate acknowledged the "arduous" task faced by its staff and examiners and noted with pride "the issue of the December exam results before Easter in spite of difficulties arising from circumstances beyond the Syndicate's control".

2. Costs and cost-cutting

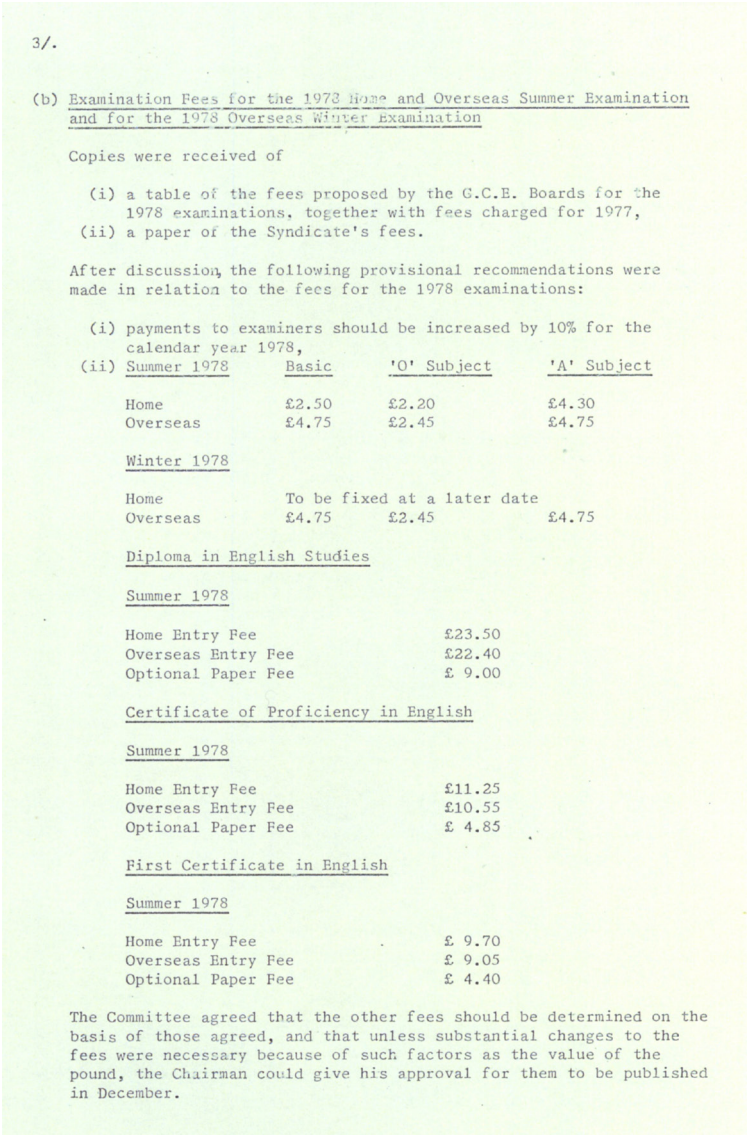
Crises are generally expensive and these disruptions to the examinations were no exception. In October 1915, the Cambridge Syndicate cited "disturbed financial conditions" to refuse a payment to a College, but cost-cutting measures were taken at

administrative level too; in 1917 it took a decision not to publish the examiners' reports, discussed sourcing cheaper envelopes, and began recording minutes on both sides of each page.

The issue of vouchers (which could be used to pay future examination fees) to centres where examinations had not taken place was normal practice, so references to refunds indicate unusual times; Asian and African centres received refunds due to the war in 1917, but the offer of refunds was clearly a contentious issue and sparked a furore when considered for UK centres affected by the flu pandemic. In the end, centres at Worthing and Petersfield received refunds while the centres at St Martin's in Scarborough and Barnard Castle were issued with vouchers, with the caveat that the Cambridge Syndicate was "open to more generous terms if [the] school [was] in difficulty".

The Cambridge Syndicate discussed issuing vouchers in lieu of examination fees for evacuees in May 1940, for use against the full amount for entry in December. For all their planning measures, it was still a period of great uncertainty, and the decision not to go ahead would have undoubtedly caused relief as the war dragged on.

It is a commonly held view that the early 1970s were lean times for the Cambridge Syndicate; Dr Frank Wild became Secretary in 1972 and at his first General Purposes Meeting, he introduced extensive cost-cutting measures, including accounting procedures and sales control, while a capital expenditure budget was also established under his watch. How much of this may have been influenced by raging inflation is unclear, but records for this period are testament to the rising costs; examination fees for UK and overseas candidates rose six times between 1973 and 1977, and in 1978 minutes show that fees paid to examiners were to be increased by 15 per cent from the following year.



3/.

(b) Examination Fees for the 1978 Home and Overseas Summer Examination and for the 1978 Overseas Winter Examination

Copies were received of

(i) a table of the fees proposed by the G.C.E. Boards for the 1978 examinations, together with fees charged for 1977,
(ii) a paper of the Syndicate's fees.

After discussion, the following provisional recommendations were made in relation to the fees for the 1978 examinations:

(i) payments to examiners should be increased by 10% for the calendar year 1978,

(ii)

Summer 1978	Basic	'O' Subject	'A' Subject
Home	£2.50	£2.20	£4.30
Overseas	£4.75	£2.45	£4.75
Winter 1978			
Home	To be fixed at a later date		
Overseas	£4.75	£2.45	£4.75

Diploma in English Studies

Summer 1978

Home Entry Fee	£23.50
Overseas Entry Fee	£22.40
Optional Paper Fee	£ 9.00

Certificate of Proficiency in English

Summer 1978

Home Entry Fee	£11.25
Overseas Entry Fee	£10.55
Optional Paper Fee	£ 4.85

First Certificate in English

Summer 1978

Home Entry Fee	£ 9.70
Overseas Entry Fee	£ 9.05
Optional Paper Fee	£ 4.40

The Committee agreed that the other fees should be determined on the basis of those agreed, and that unless substantial changes to the fees were necessary because of such factors as the value of the pound, the Chairman could give his approval for them to be published in December.

Figure 3: C/FGP/2/8 Finance and General Purposes Committee minute book 1974–1980.

3. Standards

Pressure to keep the exams running was matched by pressure to maintain examination standards. To alleviate difficulties schools experienced in obtaining suitable books, "alternative syllabus" arrangements were introduced in 1916, allowing centres to submit their own choices of syllabuses and set books of "suitable and of equivalent length and difficulty" for consideration by the Cambridge Syndicate. There is evidence too that the Cambridge Syndicate reached out to support schools in 1945 when the English syllabus was adapted to include texts that schools were more easily able to obtain.

In a more co-ordinated response "to minimise the effects of hardships", a notice was issued to schools in December 1939, outlining plans to create simpler and shorter question papers, with fewer compulsory questions, all with the aim of giving candidates a wider choice of questions. Schools were invited to submit reports before the examinations, outlining conditions of work during the year and indicating subjects in which candidates might be expected to reach a certain standard. It promised that the standard adopted by the Awarding Committee would "take full account of the standard in

a normal year and … pay attention to previous results of individual schools".

In October 1945, the Cambridge Syndicate received a bundle of documents from the Colonial Office. In it were marked candidate scripts, question papers, exam regulations and candidate results compiled by Harold Cheeseman, Examinations officer, relating to the examinations held in Sime Road Internment camp, Singapore, earlier that year. "It will mean a great deal to the candidates to get this recognition", wrote Cheeseman, adding in his accompanying letter that "it will mitigate the loss involved by the war and I have used the possibility of this recognition as an incentive not only to them but to the whole school. It has not been easy to keep up interest and effort".

Cheeseman, a pre-war Examinations officer, had set up a school at Sime Road camp, recruiting teachers and examiners. The arrangements for the examinations included details of the syllabus, centres and panel of examiners. "It was decided", explained Cheeseman, "to set and mark papers strictly in accordance with what were regarded as the normal requirements of the Syndicate". The first batch of scripts, for examinations set in January 1945, had unfortunately been destroyed, but the second, set in August, had survived and were sent to the Cambridge Syndicate for inspection. Despite obvious pressure, not least from Cheeseman and the intervention of the Colonial Office, the Cambridge Syndicate decided not to issue full certificates. Instead, based on inspection of the scripts available, it decided to issue special certificates to the candidates which stated the circumstances and standards reached "in those subjects for which it had been possible to classify results".

Outside of the context of war, an appeal from the Executive Office of Head Mistresses in November 1918 to recognise "the difficulty of schools with influenza during the present school year" persuaded the Cambridge Syndicate to allow the Secretary to "shew a slight degree of leniency [in the forthcoming examinations] with regard to the bare pass". However, a subsequent dip in standards seems to have been temporary; analysing the December 1918 examination session, the Cambridge Syndicate recorded a "general falling off of standards", but the standard of examinations the following summer session was considered "not unsatisfactory".

Standards were discussed in 1948, with regard to examinations in India, and a decision was made to raise the standard of examination in Indian languages from 1950, for both ordinary and higher level examinations. Ultimately, it was suggested, the standards would be aligned to the local matriculation or school leaving examination of the regions concerned. It was also noted that, for matriculation purposes, the standard in English Literature and Mathematics of the Cambridge Examinations was lower than the average local Indian Matriculation Examination. It was proposed that Indian universities would be consulted before any action was taken.

The postal strike of 1971 coincided with a major breach of security in the previous December examinations in East Africa, involving several groups of people in the region, as well as at the Cambridge Syndicate. This combination of difficulties would, according to the Annual Report, have "far reaching repercussions on the processing of results" and subsequent power cuts caused understandable concern for potential disruption to data processing work.

4. Supporting candidates and staff

War required not only cost-cutting exercises, but financial support, and several references to a "war bonus" during the First World War hint at a reliance people placed on it; it was raised twice in 1917 with warnings from the General Purposes Committee to "spend it wisely". A similarly stern approach was taken with appeals during the pandemic, all reflecting an era when the continuation of the exams was the primary concern, above practical logistics or the impact on individuals involved.

At the outset of the Second World War, staff welfare was addressed through the issue of an air raid precaution notice instructing staff to learn to use fire extinguishers, carry gas masks and keep office windows open so that they may hear the siren.

During the emergency in India, the Advisory Committee for Overseas Examinations' India and Pakistan sub-committee introduced measures designed to assist students. Fees for transferring centres were waived for the December 1947 session, and arrangements were approved to accept a new centre for Hindu and Sikh students on the same terms as the established college from which those students had been forced to leave. Also, as a consequence of events around Indian partition, the Cambridge Syndicate considered whether an extra session would need to be held in July 1948 to accommodate those students who had been unable to take examinations the previous December. Alongside the decision that this would not be necessary, is the note that one of the Committees would "approach the authorities in India with a view to securing that there should be no consequent hardship for University entrants".

Inflation in the 1970s caused sharp rises in candidate entry fees both in the UK and overseas throughout the decade, but there were also increases in examiner fees. The examiners who transported scripts to collection centres during the postal strike were granted double expenses, and, after the devaluation of the pound, additional payments were made to examiners in the UK. Staff and examiners who worked "exceptional hours" were warmly thanked in the 1971 Annual Report, and an acknowledgment was made to staff who worked "unusual hours" to dodge the power cuts in February 1972. The Cambridge Syndicate was clearly keen to express gratitude to staff at this time of financial hardship. By 1978, finances were stabilising; not enough to allow extra recruitment, but sufficient to drop the staff tea and coffee charges from 20 pence to "a nominal sum"!

5. Aftermath and long-term / radical changes

The shortage of suitable men to run the examinations after the First World War had a direct impact on the role of women. Not only did the regulations swiftly change to allow boys and girls to be examined in the same room, but an unprecedented motion to appoint a female presiding examiner was passed at a centre in 1916.

The most significant post-war change was the introduction of the first national examinations in 1917, when the School Certificate and Higher School Certificate replaced the old Cambridge Locals. The exams were similar in format to the Junior and Senior Locals, as certificate-based exams requiring candidates to take a range of subjects, but the introduction of these new Board of Education examinations marked the first step towards regulation of all UK school examinations.

Lack of references to the flu pandemic in the Annual Reports or Regulations could be because of its proximity to the First World War, but also due to its creeping prevalence; it was just too obvious to mention at the time. Certainly, although the impact on the exams was significant, it was relatively brief; contemporary references to "broken work" indicate an expectation of short-term disruption which, in the context of the First World War, seems not unreasonable. The records that do exist expose disunity in the Cambridge Syndicate, but the ensuing delay and inaction paid off and "normal" exam operations resumed in 1919.

At a meeting on 4 May 1944, the Cambridge Syndicate issued a statement to the President of the Board of Education: "The Syndicate consider that, before any proposal is adopted for the eventual abolition of the system of external examinations, further inquiry should be conducted into the need for so great a change". In the event, external exams were not abolished, but entered a new phase, with the introduction of the first single subject General Certificate of Education qualification six years after the end of the Second World War. But such a radical consideration, during a period of extreme upheaval, must have been deeply unsettling.

The 1941 Memorandum of Understanding with the British Council is largely credited with allowing the Cambridge Syndicate's English examinations to continue to be held at British Council premises throughout the Second World War. In 1939, there were 31 centres in Europe offering Cambridge English Exams and by 1946, despite the war, this had increased to 39. The Certificate of Proficiency in English was joined by the Lower Certificate in 1939 and the Diploma in English Studies in 1945 and candidate numbers swelled, thanks largely to entries from allied forces, including Italian and Polish prisoners of war. From just 199 candidates in 1938, the Cambridge Syndicate registered 11,258 candidates for Cambridge English Examinations in 1948.

The First Meeting of the Awarding Committee for Overseas Examinations was held on 26 February 1946; by then, Cambridge Local Examinations had been set for candidates overseas for 83 years and the first meetings dealt with several applications relating to examinations which had been held in internment camps during the war. But this newly constituted committee represented a formal move towards managing overseas examinations differently. Within 10 years it had become the Overseas Awarding Committee, from which the Council for International Examinations emerged, before the creation of an entirely separate business unit, what was to become Cambridge Assessment International Education, in 1998. This body oversaw the emergence of Local Regional Committees in the 1950s, localisation programmes in the 1960s, the development of regional departments in the 1980s and 1990s, and the establishment of Regional Offices in the 21st century.

The 1970s were testing times for the organisation—the Annual Report of 1976 references 57 theoretical studies of syllabuses and schemes of examination for proposed new examinations in "the fever for change". There is no doubt that the organisation was affected by the strikes and inflation, but it is unclear how much impact the hardships of the 1970s had on the radical transformation of school examinations in the 1980s; the rationalisation of examination boards to form regional groups to deliver the new GCSE led, ultimately, to the loss of around 19 of the examination boards which had operated in the 1970s.

Summary and discussion

Investigation of the records around the events and emergencies detailed in our archives collections reveals 10 key areas where, as an organisation, we have previous experience of similar issues and concerns to those we have seen during the 2020 COVID-19 pandemic.

1. Interruption to the practical arrangements for traditional examinations.
2. Closure of centres.
3. Missed examinations.
4. Financial disruption.
5. Loss or delay to materials.
6. New measures brought in at short notice.
7. Retuning of corporate principles or standards.
8. Loss of staff from the offices.
9. General exhaustion and distress among our staff and those we work with.
10. Stoicism under duress.

Looking at these through the lens of 2020, and broadly following the chronology of events as they unfolded in 2020, it is possible to see that our actions now are frequently in line with our actions in previous times. The examples show that in the long term, disruption to exams was no bar to significant progress in education or to achievements of the individuals affected.

History has not judged our organisation badly—we have thrived through good times and endured through bad. The crucial question of 2021, however, is not how individual organisations will fare under the current climate but how individual students may be supported to progress through further education and employment in the enforced absence and adaptation of traditional high-stakes assessments.

The sudden and immediate threat posed in March 2020 to the normal procedures and processes of the education system, and in our case especially the assessment industry, mirrors the effects seen during both World Wars and Indian Partition. Within a few weeks, the usual arrangements for assessment in centres were in sudden disarray, with practical issues such as co-ordination of pupils and the running of assessments in centres made very problematic. While centre buildings were not being physically destroyed, their partial closure had a similar effect on the viability of assessments. With the decision to move to Centre Assessed Grades, the financial costs to the organisation became apparent, in the same way as they did during the flu pandemic.

At the same time, in March 2020, closure of office buildings worldwide presented a challenge from a different angle. Processes had to be adjusted in order to work from home offices or, where that was totally impossible, safely within the office spaces. As happened during the Second World War, staff adjusted quickly and efficiently to new measures imposed to heighten their safety. Perhaps there is not that much difference between the obligatory wearing of face masks and the carrying of gas masks.

As the effects of lockdown and school disruptions continued throughout the summer of 2020, thoughts turned to the thorny issue of standards and how to apply them in such

different circumstances. Fairness to candidates has always been an underlying principle observed by the organisation, and the measures under consideration mirror very closely those that have been documented in the past—alternative syllabuses, accessible texts, simpler and shorter question papers, even the awarding of extra marks (rather appealingly called a "war bonus for exams" in Malaya in 1945) is recorded, offering wider choice of questions and leniency in applying the usual standard with regard to key boundaries. These are all very familiar areas of debate and concern in 2020 and 2021.

It has been reassuring to see that ill-tempered or discordant discussions are probably entirely normal in unusual situations when many of the protagonists are stressed across different areas of their lives. An anxiety to promote a particular view of the best solution to a problem during a pressured situation can play havoc with social niceties, as evidenced by minutes and notes from the past. While it is tempting to suggest that the sheer numbers of competing voices might have been smaller in the past, the size of committees (for example, those dealing with Indian Partition) are similar to the size of committees today.

It is interesting to contrast the approach to long- and short-term stress for both staff members and key stakeholders such as students and teachers, with the past. Endurance seems to have been the watchword for previous generations and has, certainly, been a feature of the present situation. However, in the 21st century there is positive encouragement to voice anxieties and a great deal more effort is taken to protect, as best it can be, the mental health of all concerned. Previous generations' stoicism (carrying on regardless, even when under extreme duress) has been replaced by a more understanding approach.

The other key contrast we can see with past events is the benefit of modern technology, which has allowed us to solve problems in ways that were simply not available to our predecessors. However, there are similarities with the fact that changes have been implemented that may prove enduring.

At the time of writing, the COVID-19 pandemic is far from over and new disruptions may be ahead; these extracts from the archives are not exhaustive and, in some cases, there is a frustrating lack of detail to draw on, but they highlight some of the difficulties, some of the measures put in place to deal with them, and some of the outcomes. Although our study has been limited to a few key historical events, it has become apparent that almost all of the measures discussed in 2020 have been considered, or implemented, in the past.

References

Publications

Cooke, G. (2020, May 6). How exams survived the war. *Cambridge Assessment Website Blog.* https://www.cambridgeassessment.org.uk/blogs/ve-day-how-exams-survived-the-war/

Elliott, G. (2011). 100 years of controversy over standards: an enduring problem. *Research Matters: A Cambridge Assessment publication, Special Issue 2,* 3–8. https://www.cambridgeassessment.org.uk/Images/567572-100-years-of-controversy-over-standards-an-enduring-problem.pdf

Newton, P. (2011). A Level pass rates and the enduring myth of norm-referencing. *Research Matters: A Cambridge Assessment Publication, Special Issue 2*, 20–26. https://www.cambridgeassessment.org.uk/Images/567580-a-level-pass-rates-and-the-enduring-myth-of-norm-referencing.pdf

Raban, S. (Ed.). (2008). *Examining the World. A History of the University of Cambridge Local Examinations Syndicate.* Cambridge University Press.

Cambridge University Archives

LES 1/6; Syndicate Minutes October 1914, May & October 1915.

LES 1/7; Syndicate Minutes May 1916 & November 1918.

Cambridge Assessment Archives

Printed matter

Annual Accounts 1970 to 1979.

Annual Reports 1914 to 1919, 1939–1945, 1970–1976.

Bound Volumes of Examination Regulations 1914, 1916, 1939, 1940, 1947.

Class list 1944.

Joint Committee of the Local Exams Syndicate and the British Council Cambridge

Examinations in English Survey for 1948.

Unprinted sources

A/F 3/4; Finance Administration, Clerks Wages Book, 1892–1919.

A/WW11 1/4; Certificates of Supervision and Correspondence on war difficulties, 1943–1944.

A/WW11 1/3; Papers referring to internment camp examinations in Singapore, 1944–1945.

C/CB 1/2 (Syndicate); May & June 1917, October 1918, January 1919 & October 1919.

C/CB 1/2 (General Purposes Committee); June & October 1917.

C/CB 1/2 (Joint Committee on Examinations); November 1918.

C/FGP 2/6; Finance and General Purposes Committee; February 1973 & July 1978.

C/FGP 2/8; Finance and General Purposes Committee 1974–1980.

EX/S 1/11; Syndicate Minutes October 1939, January 1941 & October 1945.

EX/S 1/16; Syndicate Minutes, January 1971; General Purposes Minutes February 1972.

EX/S 1/17; Syndicate Minutes, November 1973.

M/REG; Scrapbook of Regulations, June, September & December 1939.

O/COE 2/1; Council for Overseas Exams, November 1945.

O/IPAC 2/1; India and Pakistan Regional Advisory Committee, 1947–1962.

O/MAC 2/1; Malaya Advisory Committee, February 1946.

Cambridge Assessment

Postgraduate Advanced Certificate in *Educational Assessment*

Stand out as an assessment expert

Our University of Cambridge Postgraduate Advanced Certificate in Educational Studies: Educational Assessment (PGCA) is a practice-based qualification. The course is taught part-time over 15 months, giving you 90 credits at Master's level (Level 7).

You'll learn through a mix of online learning and four day schools led by experts from Cambridge Assessment and the University of Cambridge Faculty of Education.

On successful completion of the course you'll be awarded a Postgraduate Advanced Certificate in Educational Studies (PACES).

Take our course to:

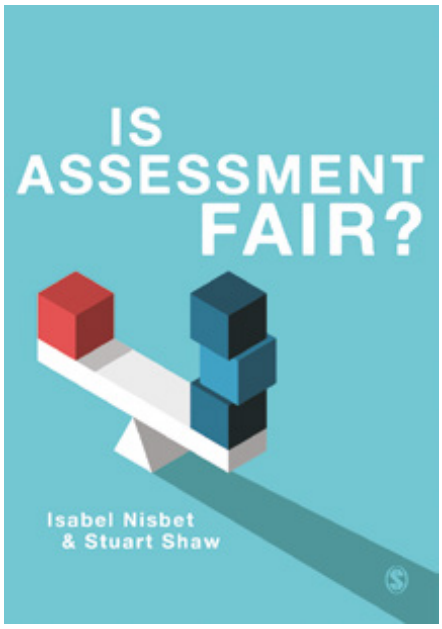
- develop an in-depth understanding of assessment in a supportive academic environment
- work on small-scale enquiries relevant to your own professional practice
- apply research methodologies to your professional context
- learn from other students and build a network beyond your specialist area.

Starting September 2021

cambridgeassessment.org.uk/pgca

Is Assessment Fair?

Isabel Nisbet -
University of Cambridge
Stuart Shaw -
Cambridge Assessment



Fairness in educational assessment has become a major talking point and allegations that assessments are unfair are commonplace on social media and in the press. But what does fairness mean in practice and how can we evaluate it?

This book offers a timely and necessary investigation, exploring the concept through the lenses of: measurement theory, social justice, the law and philosophy in order to put forward a template for fairness in educational assessment.

Drawing on international examples from the UK, US, Australia and South East Asia, this book offers a commentary on fairness that is highly relevant to the changing context of assessment today.

This book will be of interest to anyone with a professional or academic interest in educational assessment, to education policymakers and to all who are working to make assessment fair.

Table of Contents

Everyone wants educational assessment to be valid and fair, but what do these terms really mean in practice? Over the last 50 years, a great deal of effort has gone in understanding what we really mean by validity, but what we mean by fairness has received little attention – until now. In this important, timely, and highly readable book, Isabel Nisbet and Stuart Shaw offer us a comprehensive set of perspectives on assessment – theoretical, professional, legal, philosophical and social justice – that enables to see how debates about whether particular assessment practices are fair are often really disguised debates about what fairness means. As issues of fairness assume greater importance in the coming years, the book will be an invaluable guide to thinking clearly about the challenges that all those professionally and personally involved in educational assessment will face in making assessment fairer.

Prof. Dylan Wiliam, Emeritus Professor of Educational Assessment, Institute of Education, University College London

Nisbet and Shaw provide a comprehensive and thoughtful discussion of fairness in assessment, and far beyond assessment, employing multiple lenses (measurement theory, professional standards, law, philosophy, and social justice) to examine fairness from different angles, and their analyses of the concept will certainly enrich discussions of its use in assessment.

Prof. Michael Kane, ETS

This is a wide-ranging – from Aristotle to Big data – view of fairness in educational assessment, that covers usually neglected areas such as philosophical and juridical underpinnings, while carefully explaining technical aspects in accessible terms. Scholarly and thorough, it never loses sight of the fact that fairness affects real people – teachers, parents, employers, higher education and most importantly students – who sometimes have only one shot at doing well in an increasingly narrow assessment system.

Dr. Tina Issacs, Honorary Associate Professor, UCL Institute of Education

Although the assessment field has been engaged with the societal issues of fairness, they have often been caught unawares and we have found ourselves fighting fires rather than at the vanguard. This book is different, in that it raises broad, societal issues connected with assessment fairness. It does so in a scholarly manner and recognises the technical solutions that are available without constraining the discussion to them. Key strengths of the book are the multi-disciplinary approaches taken and the insightful selection of examples, which will become standard reading in many courses. This book raises issues about systemic unfairness in ways that have not been seen for some time; they are modernised here, through cases, references to empirical work and in the theoretical framing. The book is thought-provoking, will stimulate debate and hopefully will fire up the field to connect with the big issues it raises.

Prof. Jo-Anne Baird, University of Oxford

Sept 2020
Paperback £24.99 (9781526474926)
Hardback £75.00 (9781526474940)

Research News

Anouk Peigne Research Division

Tim Oates talks in his Foreword to this edition of *Research Matters* about the impact of the pandemic on our work at Cambridge Assessment. Another very noticeable effect has been the sharp rise in the volume of discussion in the specialist and mainstream domestic media about aspects of the UK education system. Much commentary has, in the past, tended to focus on expunging GCSEs, with calls tending to come from the same few individuals, often motivated by limited self-interest. However, the events of the past 12 months have catalysed opinion on this matter and others. Often assertions have been made without supportive facts. It is right to debate such issues, but we believe in setting our views in the context of evidence.

Cambridge Assessment and Cambridge University Press have chosen to contribute to the debate initially through a series of blogs, research papers and events set around 12 outline principles for the future of education (see page 106). Over the course of this year, these will build into a rich compendium of thought, but also clear prescription for action. To date, we have written about how external assessment at the end of basic secondary education is more common around the world than is often suggested, how hurricanes and earthquakes can inform recovery learning in the wake of the current crisis, how education systems are adapting to the increased use of technology, and how assessment systems might be made more resilient. You can find these articles and more on our website.

Chris Shadforth

Director of Communications

Blogs

The following blogs have been published since *Research Matters*, Issue 30:

Oates, T. (2021, March 4). *What hurricanes and earthquakes can tell us about Coronavirus 'recovery learning'.* https://www.cambridgeassessment.org.uk/blogs/coronavirus-recovery-learning/

Oates, T. (2021, March 9). *Professor Jane Mellanby 1938-2021.* https://www.cambridgeassessment.org.uk/blogs/professor-jane-mellanby-obituary/

Oates, T. (2021, March 18). *The importance of the individual – now, more than ever.* https://www.cambridgeassessment.org.uk/blogs/the-importance-of-the-individual/

Suto, I. (2021, February 24). *Exams at 16? They're more common around the world than perhaps you thought…* https://www.cambridgeassessment.org.uk/blogs/exams-at-16-more-common-around-the-world-than-you-thought/

Conference Presentations

Education Assessment Research Seminar

In October 2020, Cambridge Assessment organised a series of virtual themed sessions, reflecting Ofqual's April 2020 seminar which had to be cancelled due to COVID-19. The themes were vocational qualifications, accessibility, marking and maintaining standards. Researchers of Cambridge Assessment presented various papers:

Tom Benton: *Introducing the simplified pairs method for standard maintaining.*

Vicki Crisp and Sylwia Macinska: *Creating better tests—students' views on the accessibility of different question design features.*

Sarah Hughes and Tony Leech: *OCR trials of comparative judgement for awarding: reliability, utility and validity considerations.*

Sylwia Macinska and Carla Pastorino: *Embedded tools for accessible on-screen assessment: a review of the evidence on effectiveness.*

Stuart Shaw and Isabel Nisbet: *What makes vocational assessments fair?*

Antonia Sudkaemper and Breanne Chryst: *Review of marking—comparing different types of feedback.*

Carmen Vidal Rodeiro and Sylwia Macinska: *Impact of access arrangements on performance.*

Sylvia Vitello and Carmen Vidal Rodeiro: *Vocational qualifications at Key Stage 4 and Key Stage 5: who takes them and how they fit into students' programmes of study.*

Emma Walland: *Leadership in the context of examining: perspectives of assessment specialists.*

The sessions are available to watch on the Cambridge Assessment website: https://www.cambridgeassessment.org.uk/news/global-audience-for-assessment-research-seminar/

AEA-Europe Online Conference

As Chair of the Scientific Programme Committee (SPC), Stuart Shaw from the Cambridge Assessment International Education organised the November 'Festival' which was a distilled version of the normal conference.

He also participated in various sessions: he acted as a panellist for *Assessment Cultures SIG. Focus: How did your education and assessment system react to the effects of Covid-19?* He co-presented *Life and Times of the Scientific Programme Committee: the work and remit of the SPC* with Elisa de Padua and Nico Dieteren. He co-chaired a PhD Session with Christina Wikström. Finally, he participated in the General Assembly *Presentation of the Annual Report on behalf of the Scientific Programme Committee.*

International Conference of Education, Research and Innovation

The annual ICERI conference took place online in November. In 2020, this conference covered various topics including Distance Learning in Times of Crisis, Pedagogical Methods and Innovations, Global Issues in Education and Research and brought together

participants from more than 70 countries.

Emma Walland and Stuart Shaw presented *Using e-portfolios to capture and assess transversal skills: tensions in theory and praxis.*

Publications

The following reports and articles have been published since *Research Matters*, Issue 30:

Benton, T., Cunningham, E., Hughes, S., & Leech, T. (2020). *Comparing the simplified pairs method of standard maintaining to statistical equating*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. https://www.cambridgeassessment.org.uk/Images/599225-comparing-the-simplified-pairs-method-of-standard-maintaining-to-statistical-equating.pdf

Benton, T., Leech, T., & Hughes, S. (2020). *Does comparative judgement of scripts provide an effective means of maintaining standards in mathematics?* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. https://www.cambridgeassessment.org.uk/Images/603922-does-comparative-judgement-of-scripts-provide-an-effective-means-of-maintaining-standards-in-mathematics-.pdf

Cambridge Assessment. (2020). *Consultation regarding a potential GCSE in Natural History: Early findings.* Cambridge, UK: Cambridge Assessment. https://f.hubspotusercontent20.net/hubfs/4123326/Natural%20History%20Files/Natural%20History%20Consultation%20Report_110920.pdf?__hssc=251199705.7.1605626740691&__hstc=251199705.

Constantinou, F. (2020). What is synoptic assessment? Defining and operationalising an as yet non-mainstream assessment concept. *Assessment in Education: Principles, Policy and Practice, 27*(6), 670–686. https://doi.org/10.1080/0969594X.2020.1841734

Crisp, V., Shaw, S. D., & Bramley, T. (2020). Should we be banking on it? Exploring potential issues in the use of 'item' banking with structured examination questions. *Assessment in Education: Principles, Policy & Practice, 27*(6), 655–669. https://doi.org/10.1080/0969594X.2020.1827220

Fitszimons, S., & Johnson, M. (2020). How Collaborative Project Development Theory Can Be Used to Provide Guidance for International Curriculum Partnerships. *International Dialogues on Education: Past and Present, 7*(2), 24–39. https://www.ide-journal.org/article/2020-volume-7-number-2-how-collaborative-project-development-theory-can-be-used-to-provide-guidance-for-international-curriculum-partnerships/

Gill, T. (2020). *The relationship between taking a formal music qualification and overall attainment at Key Stage 4.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. https://www.cambridgeassessment.org.uk/Images/603850-the-relationship-between-taking-a-formal-music-qualification-and-overall-attainment-at-key-stage-4.pdf

Gill, T. (2020). *An analysis of grade boundary marks in A levels and GCSEs if the 2020 'standard' was applied to 2019 results.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. https://www.cambridgeassessment.org.uk/

Images/604501-an-analysis-of-grade-boundary-marks-in-a-levels-and-gcses-if-the-2020-standard-was-applied-to-2019-results.pdf

Nisbet, I., & Shaw, S. D. (2020). *Is Assessment Fair?* Sage. https://uk.sagepub.com/en-gb/eur/is-assessment-fair/book266191

Shaw, S. D. (2020). Achieving in Content Through Language: Towards a CEFR Descriptor Scale for Academic Language Proficiency. In M. deBoer, & D. Leontjev (Eds.), *Assessment and Learning in Content and Language Integrated Learning (CLIL) Classrooms: Approaches and Conceptualisations.* (pp.29–56). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-54128-6_2

Suto, I., & Oates, T. (2021). *High-stakes testing after basic secondary education: How and why is it done in high performing education systems?* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. https://www.cambridgeassessment.org.uk/Images/610965-high-stakes-testing-after-basic-secondary-education-how-and-why-is-it-done-in-high-performing-education-systems-.pdf

Data Bytes

Data Bytes is a series of data graphics from Cambridge Assessment's Research Division, designed to bring the latest trends and research in educational assessment to a wide audience. Topics are often chosen to coincide with contemporary news or recent Cambridge Assessment research outputs. The following Data Byte has been published since Research Matters, Issue 30:

- December 2020: Does comparative judgement provide an effective means of maintaining standards in mathematics?

Sharing our research

We aim to make our research as widely available as possible. Listed below are links to the places where you can find our research online:

Journal papers and book chapters: www.cambridgeassessment.org.uk/our-research/all-published-resources/journal-papers-and-book-chapters/

Research Matters (in full and as PDFs of individual articles): https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

Conference papers: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/conference-papers/

Research reports: www.cambridgeassessment.org.uk/our-research/all-published-resources/research-reports/

Data Bytes: www.cambridgeassessment.org.uk/our-research/data-bytes

Statistics reports: https://www.cambridgeassessment.org.uk/our-research/all-published resources/statistical-reports/

Blogs: www.cambridgeassessment.org.uk/blogs/

Insights (a platform for sharing our views and research on the big education topics that impact assessment around the globe): https://www.cambridgeassessment.org.uk/insights/

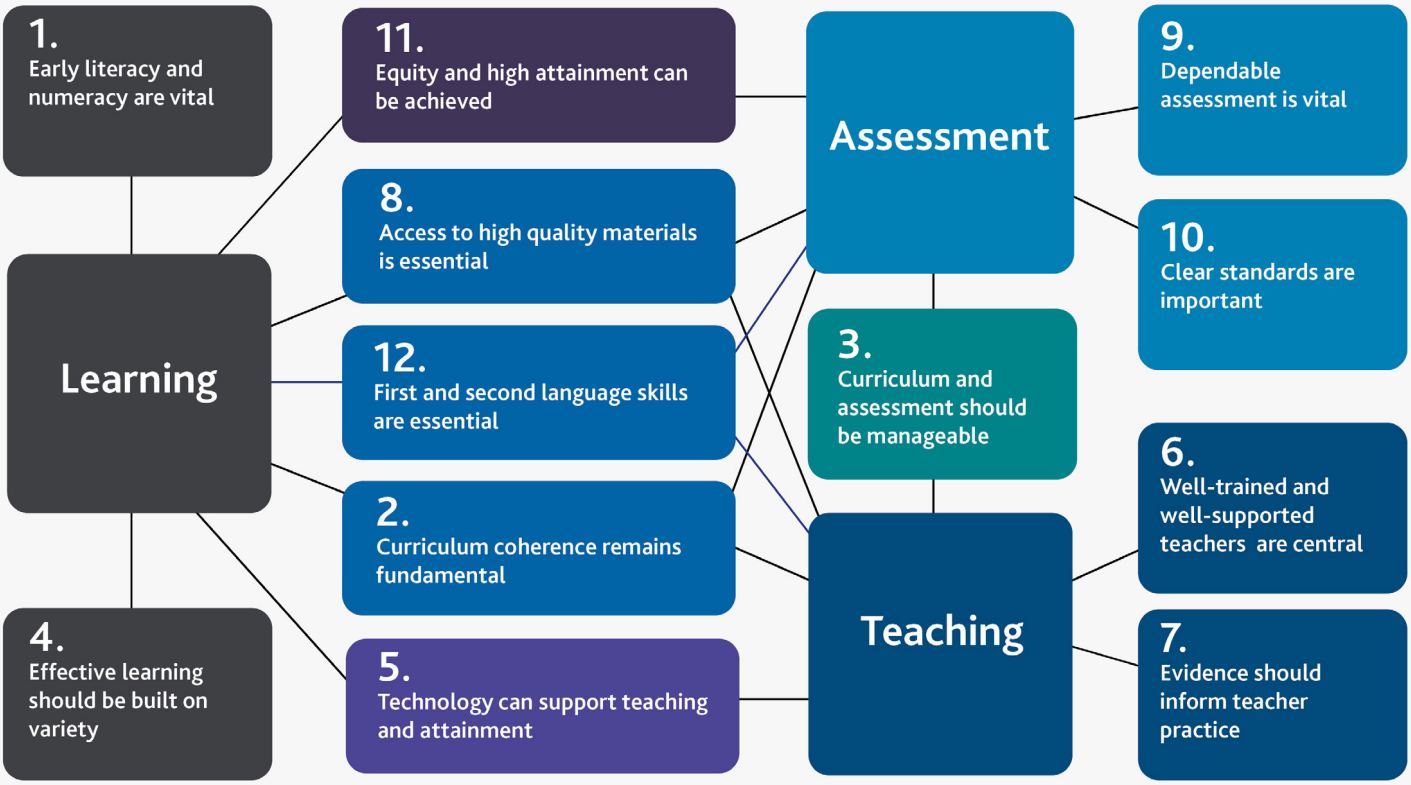
Our YouTube channel: www.youtube.com/user/CambridgeAssessment1 contains Research Bytes (short presentations and commentary based on recent conference presentations), our online live debates #CamEdLive, and podcasts.

You can also learn more about our recent activities from Facebook, Instagram, LinkedIn and Twitter.

Outline principles for the future of education

We have published 12 outline principles for the future of education, as a mechanism for navigating debates around the future of teaching, learning and assessment. We are using this approach of interconnected principles to help focus and refine thinking and call out challenges where we see them.

Outline principles for the future of education



Irenka Suto and Tim Oates CBE from our Research Division have also released a new report looking at the use of high stakes assessment at age 15/16 in repeatedly high performing jurisdictions. They find that exams at 16 are more common around the world than you may think.

To find out more read our research and blogs cambridgeassessment.org.uk/future-of-education

Cambridge Assessment Network

Take your assessment ambitions further with our range of online CPD accredited courses. With multiple start dates throughout the year, our 9-week online courses can be taken at a time to suit you.

A101: Introducing the Principles of Assessment

Understand what's behind good assessment with a detailed grounding in the principles of assessment.

A102: Introducing Assessment Practice

Learn how exams are created and marked as you discover the key stages of the assessment cycle.

A103: Introducing Data Literacy

Make greater use of assessment data and understand the basics of managing, reading and interpreting data.

A104: Psychometrics in Educational Assessment

Expand what you know about assessment analysis in order to ensure tests are reliable, valid and fair.

Find out more at: cambridgeassessment.org.uk/online

Contents / *Issue 31* / Spring 2021

Cambridge Assessment
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

+44(0)1223 553985
researchprogrammes@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk