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