What are our key assessment principles after a summer without assessment?

Richard Harry

A Research & Development Project

Submitted for the MSc Educational Assessment 2021

DEPOSIT AND CONSULTATION OF THESIS

One copy of your dissertation will be deposited in the Department of Education Library via Oxford Research Archives where it is intended to be available for consultation by all Library users. In order to facilitate this, the following form should be inserted in the library copy of the dissertation.

Note that some graphs/tables may be removed in order to comply with copyright restrictions.

Surname	Harry
First Name	Richard
Faculty Board	Education
Title of Dissertation	What are our key assessment principles after a summer without assessment

Declaration by the candidate as author of the dissertation

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.

2. I understand that the Department requires that I shall deposit a copy of my dissertation in the Department of Education Library via Oxford Research Archives where it shall be available for consultation, and that reproductions of it may be made for other Libraries so that it can be available to those who to consult it elsewhere. I understand that the Library, before allowing my dissertation to be consulted either in the original or in reproduced form, will require each person wishing to consult it to sign a declaration that he or she recognises that the copyright of this thesis belongs to me. I permit limited copying of my dissertation by individuals (no more than 5% or one chapter) for personal research use. No quotation from it and no information derived from it may be published without my prior written consent and I undertake to supply a current address to the Library so this consent can be sought.

3. I agree that my dissertation shall be available for consultation in accordance with paragraph 2 above.

What are our key assessment principles after a summer without assessment?

Abstract

The Covid-19 pandemic caused the cancellation of 2020 public general qualification examinations in the United Kingdom (UK), and elsewhere. The implementation and subsequent rejection of policies of grade standardisation that followed across the UK's four nations has the potential to have long-lasting effects on the UK's qualification systems. The aim of this study was to explore the short- and long-term effects of the policies relating to the summer 2020 general qualification examination series, with respect to the key assessment principles of validity, reliability, fairness, and comparability. Validity arguments for the different approaches taken were developed, using a phased approach based on Crooks, Kane and Cohen's (1996) validity chain model.

Focused on Wales as a main 'home-international' case study, while drawing on other UK nations for elucidation and comparison, this study found that, as well as the validity argument for standardisation, there were also two distinct validity arguments for the awarding of centre assessment grades (CAGs) determined by teachers. Both arguments de-prioritised comparability between examination series and centres, previous students' legitimate expectations that their grades would continue to represent an equivalent level of attainment to the grades awarded in future, and the relational sense of fairness.

The first validity argument for CAGs was based on the impossibility of meeting the twin goals of awarding fair grades to individual students and achieving causal, cohort-level comparability of outcomes, given that teachers were known from previous years to be optimistic in their predicted grades. Analysis of grade data also suggested that teachers were faced with selecting a grade when, in many cases, a multitude of grades were possible, but none very likely. Student-level standards were prioritised, thus rejecting the logic of the standardisation model. As such, the first argument

represented a temporary re-prioritisation of key assessment principles to give more weight to CAGs, as the main source of student-level attainment data.

The second validity argument for CAGs is a more fundamental departure from the usual balance of principles, stating that teachers are not so much more *optimistic* as more *knowledgeable* about a student's level of attainment, which the normal general qualifications assessment process also fails to capture, either because some students will have a bad day, or because outcomes are unfairly capped as a result of predictions used to ensure statistical comparability (Casella et al., 2021). Thus, the attainment measure from previous years was presented as insufficiently valid, and the notion of maintaining the attainment standard was rejected.

The validity arguments are then considered in relation to paradigms of assessment (Baird, 2018a) and paradigm change (Isaacs & Gorgen, 2018) to examine how the pandemic may affect proposals for future assessments and the balance of key principles they represent.

Table of contents

Section 1: Introduction

The United Kingdom's school examination system

Education policy is a devolved matter in the United Kingdom (UK). The government of each of the four nations of the UK – England, Scotland, Wales, and Northern Ireland – determines its own approach to qualifications (Raffe, 2006; Ofqual, Qualifications Wales & CCEA Regulator, 2019a). The structure, roles, and responsibilities of the various bodies responsible for determining assessment principles, designing and delivering qualifications for school-aged learners also differ.

Some common features do exist, reflecting historical alignment in qualifications prior to devolution. England, Wales, and Northern Ireland have three main types of 'general qualification': GCSEs, AS, and A-levels. These are high-stakes examinations taken by most learners in a range of subjects and are a crucial element of learners' decisions relating to further study or employment, and of their wider educational experience (Ofqual, 2019a). GCSEs are predominantly entered by 16-year-olds at the end of secondary-level schooling (He et al., 2015). Just under half of learners then progress to GCE qualifications (Ofqual, 2021): AS qualifications are predominantly taken by 17-year-olds, and A-levels are predominantly taken by 18-year-olds (Ofqual, 2020a; Qualifications Wales, 2020a). A-levels are used as a basis for selection to university courses (Newton, 2011), as well as a basis for progression to other forms of study. Each nation has its own suite of general qualifications, overseen by a regulator with statutory powers with responsibility for its nation's qualifications system – Ofqual (in England), Council for the Curriculum, Examinations & Assessment (CCEA) Regulation in Northern Ireland, and Qualifications Wales. Each regulator is operationally independent, and in England and Wales report to their nation's legislature via standing scrutiny committees. Independence is not absolute, however: the regulators' board members are appointed by their nations' respective governments, who may also give directions to their regulator in specific circumstances (Ofqual, 2019b; Qualifications Wales, 2020b; CCEA, 2020a).

Ofqual oversee a market for the awarding of general qualifications in England, within which four exam boards operate: AQA, OCR, Pearson, and WJEC (operating under the Eduqas brand) (Ofqual, 2020b). In Wales, WJEC are the sole provider of Qualifications Wales-regulated general qualifications, although a small number of Ofqual-regulated general qualifications are available where no qualification has been developed specifically for Wales (Qualifications Wales, 2020c). The awarding organisation part of CCEA are the sole provider of Northern Ireland-specific general qualifications, although many general qualifications regulated by Ofqual and Qualifications Wales can be offered by schools and colleges in Northern Ireland (CCEA, 2020b).

As the types of general qualification are common across the three nations, comparability of standards in these qualifications is a key consideration in each examination series (Ofqual, Qualifications Wales & CCEA Regulator, 2019a; Qualifications Wales, 2017), despite some divergence in content, assessment and qualification structure. At GCSE, Ofqual-regulated qualifications are awarded using a 9-1 grade scale. In Wales, an A*-G scale is used, and a modified A*-G scale partially aligned to the 9-1 scale is used in Northern Ireland (Ofqual, Qualifications Wales & CCEA Regulator, 2019b). The grade scales for AS (A-E) and A-level (A*-E) are common to all three nations. Many learners complete vocational and technical qualifications (VTQs) alongside or instead of general qualifications. Such qualifications are often offered on the same basis in England, Wales, and Northern Ireland, meaning that regulators must agree on a common position when issues arise (*cf.* Qualifications Wales, 2020d). In some sectoral areas, nation-specific VTQs are now being developed (*cf*. WJEC and City & Guilds, 2020).

Scotland's qualification structures are distinct from the other UK nations. The Scottish Qualifications Authority (SQA) devises and develops its 'National Qualifications', which are broadly equivalent to the general qualifications of England, Wales and Northern Ireland and are graded A-D. 'National 5' qualifications are predominantly taken by S4 learners (aged 15) and are designed to lead to further study, apprenticeships or employment. Higher qualifications (mostly taken by S5 learners, aged 16)

and Advanced Higher qualifications, mostly taken by S6 learners aged 17) are aimed at facilitating progression to further education and employment opportunities (SQA, 2014; 2019). There is no equivalent regulatory body: SQA's Board of Management is appointed by the Scottish Government, who also sign off their strategic plan (SQA, 2020a).

The summer 2020 school examination series

The emergence of the SARS-COV-2 (Covid-19) virus in late 2019, and the resulting pandemic, led to unprecedented disruption to education systems across the world. In the UK, increasing community transmission of the virus led to the closure of schools in March 2020 and, later that month, the cancellation of public examinations scheduled for the summer of that year (Welsh Government, 2020a; Department for Education, 2020a). In place of examinations, governments and policymakers sought to find alternative methods for awarding grades to learners so that they could progress to the next stage of their education, training, and/or employment.

Although each nation defined its own approach, some commonality in principles can be identified, as shown in *Table 1*. All approaches were intended to maximise the number of learners receiving an outcome; only some 'private candidates' not studying at an examination centre (a school, college or other educational provider registered to administer general qualifications) did not receive a grade. Although different terminology was used – grades were to be moderated, calculated, or standardised depending on the nation – the process was essentially the same, and is referred to hereafter as standardisation. First, information was gathered from centres on the level of attainment that they believed each candidate had achieved. In England, Wales, and Northern Ireland, this took the form of the grade that the centre thought was most likely to be achieved (an *absolute* judgement, referred to as a *centre assessment grade* (CAG)); centres were also asked to rank their learners (a *relative* judgement within their cohort of learners). In Scotland, grade estimates are collected from schools in normal exam series, in nine bands – two bands each for grades A, B and C, one band for D, and two bands for 'no award'. This was expanded to nineteen

refined bands for 2020; no rank order was collected beyond this, although bands were linked to grades, meaning that Scotland also had a form of CAGs (SQA, 2020b).

The second stage was the application of the standardisation processes, which were deployed at subject level (*cf*. Ofqual, 2020c; SQA, 2020c). For most students, CAGs were not used as the basis for grades. Two types of standardisation process were used in England, Wales, and Northern Ireland. In the *first* process type (*Standard Model* hereafter, see *Figure 1*), used for all general qualifications in England, most GCSEs in Wales, and all AS qualifications in Wales and Northern Ireland, centres were allocated a set of grades based on their historic cumulative grade distribution, adjusted for the relative strength of the centre's entry cohort in 2020 as measured by students' prior attainment in previous qualifications. These grades were then allocated to that centre's students based on the rank order provided by the centre (Ofqual, 2020c).

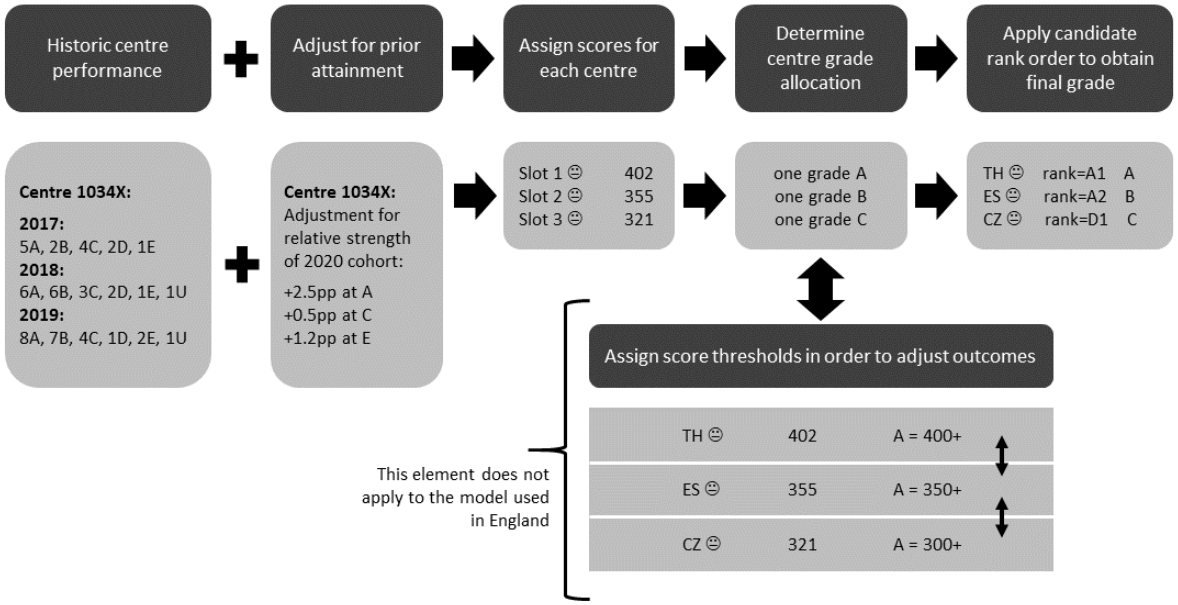
In the *second* process type (*Banked Unit Model* hereafter, see *Figure 2*), used for some GCSEs in Wales, and all A-levels in Wales and Northern Ireland, marks achieved in a unit (or units) taken previously by most students were used to predict outcomes. The grades estimated by each centre's cohort were then allocated as a set to that centre, and then distributed to students based on their centre's rank order. This model was not used in England as no units could be banked in their 'linear' qualifications prior to the summer series. In *Figures 1 and 2*, the models are illustrated for three candidates, labelled TH, ES, and CZ.

In both models, centres were allocated grades for the main age group using the described method. CAGs were used in two limited ways: to inform the outcomes of small cohorts, for whom relying on historical data was found to be unreliable in pre-award testing, in the Standard Model; and alongside the rank order data to determine grades for students not included in the main model (a process known as 'slotting in') (WJEC, 2020a; CCEA, 2020c).

Both models produced, at an intermediate stage, a set of scores which are a product of the model rather than an outcome of performances. This results in a score distribution for the whole cohort, enabling 'cut scores' to be set to produce the desired grade distribution. In Wales, all awards were subject to a review by a WJEC decision-making group, who could request an amendment to the thresholds to raise or lower subject outcomes where required (WJEC, 2020a). A similar process was employed by SQA in Scotland, although their 'mathematical optimisation' process explicitly compared each centre's estimates against an expected projection, with grades adjusted where the difference was greater than a stated tolerance (SQA, 2020b). Unlike the other models, therefore the SQA model used teacher-estimated grades as a starting point for determining grades, meaning that – unlike the CAGs in the two standardisation models – the grades had a formal meaning as a starting point for awarding for all candidates.

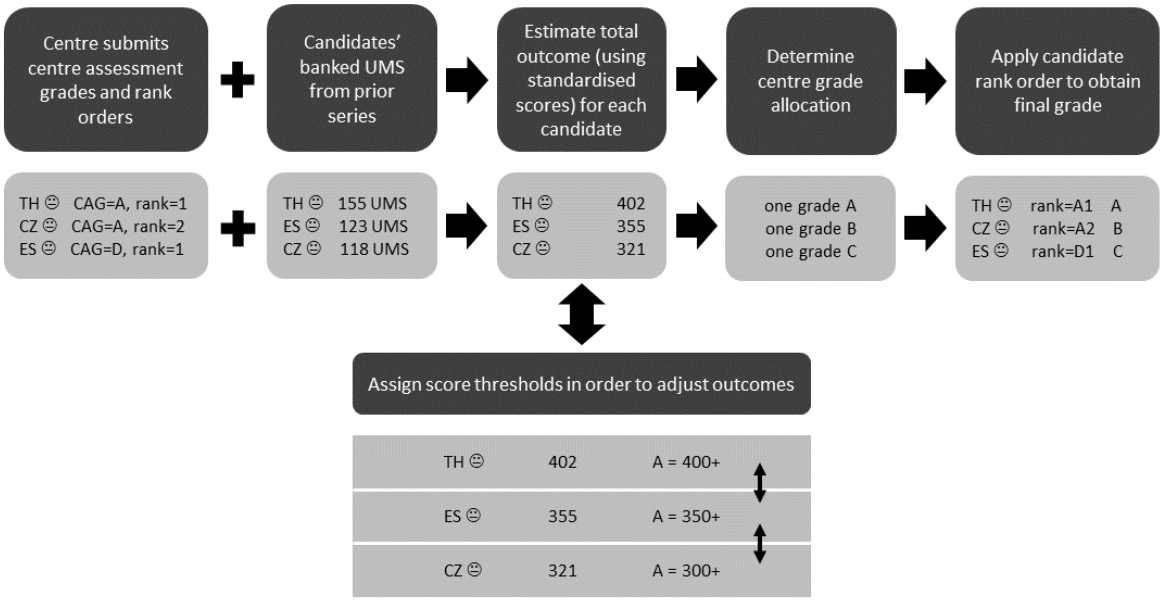
Another common feature across all four nations was the desire to maintain similar national outcomes to what would likely have occurred in summer 2020. In England and Wales (but not Scotland or Northern Ireland), the regulators consulted on their approaches – although not the standardisation process – and in both nations a majority was found to be in favour. Although some elements were the subject of some dissent, in particular the potential reliance on historical data as a basis for determining grade outcomes for centres (and, by extension, students) in 2020 (Ofqual, 2020d; Qualifications Wales, 2020e).

Figure 1: Standard Model of standardisation



Source: adapted from Ofqual (2020c), WJEC (2020b)

Figure 2: Banked Unit Model of standardisation



Source: adapted from WJEC (2020a)

Table 1: Approaches to grading for general/National qualifications in summer 2020, for the four UK nations

Nation	England	Scotland	Wales	Northern Ireland
Task for regulator, as defined by government	Calculated grades based on centres' judgements, standardised across centres, maintaining qualification standards with a similar grade profile to previous years	Moderation of teachers' judgements	Standardisation of centres' judgements	Calculate grade outcomes
Published aims/principles for the approach(es) taken	Five aims (Ofqual, 2020d): • provide students with the grades that they would most likely have achieved had they been able to complete their assessments, placing more weight on centres' historical performance (adjusted for students' prior attainment) than CAGs where this improves model accuracy; • apply a common standardisation approach, for as many students as possible; • to protect students from being systematically advantaged or disadvantaged by their socio-economic background or (protected) characteristic; • to be deliverable by exam boards in a consistent and timely way that they can quality assure and can be overseen effectively by Ofqual • use a transparent, explainable method, to encourage engagement and build confidence.	Three principles (SQA, 2020b): • fairness to all learners; • safe and secure certification of qualifications; • maintaining the integrity and credibility of the qualifications system, ensuring that standards are maintained over time, in the interest of learners.	Four aims (Qualifications Wales, 2020e): • grades for all candidates; • national outcomes to be broadly similar to those in previous years (to reduce the risk of unfairness for learners over time and maintain public confidence); • where possible, not to systematically advantage or disadvantage groups of learners protected by equalities legislation; • use a range of evidence to calculate the likely grades that learners would have achieved, had they been able to complete their assessments.	Two aims (CCEA, 2020c): • Combine teachers' judgements and statistical modelling to calculate grades. • Account for accuracy, equality and public acceptability/understanding in the calculation method.
Level at which standardisation was applied	Centre. National level (across exam boards), by subject.	Centre, by subject.	Centre, by subject; adjustments made at national level by moving model thresholds.	Centre, by subject.

Nation	England	Scotland	Wales	Northern Ireland
Data gathered from centres	Centre assessment grade (CAG), representing the grade each learner would have been most likely to achieve, and rank order position for each candidate, signed off by Head of Centre.	Centre grades (refined grade bands) and rank order position, signed off by Head of Centre.	Centre assessment grade (CAG), representing the grade each learner would have been most likely to achieve, and rank order position for each candidate, signed off by Head of Centre. *Tied ranks allowed in some circumstances.*	Centre assessment grade (CAG), representing the grade each learner would have been most likely to achieve, and rank order position for each candidate, signed off by Head of Centre.
Other data gathered for standardisation	Candidate prior attainment (to calculate centre value-added), centre historical grade profile.	Candidate prior attainment (for Higher and Advanced Higher only, not used in standardisation), centre historical grade profile. Note: candidate information was not used.	Candidate prior attainment, centre historical grade profile.	Candidate prior attainment (AS and A-level only), centre historical grade profile.
Brief description of standardisation approach	See *Figure 1*.	Mathematical optimisations; tolerances based on historical data applied to centre grades.	See *Figure 1* (most GCSEs, AS) and *Figure 2* (some GCSEs, A-level)	See *Figure 1* (GCSEs, AS) and *Figure 2* (some GCSEs, A-level) – note that no prior attainment adjustment was made for GCSE.
Derivation of final outcome	Grade awarded was the higher of: • calculated outcome • CAG	Grade awarded was the higher of: • calculated outcome • teacher estimate	Grade awarded was the higher of: • calculated outcome • CAG • AS grade achieved in 2020 or previously (for A-levels only)	Grade awarded was the higher of: • calculated outcome • CAG

Sources: CCEA (2020c), Department for Education (2020b), Ofqual (2020c,d), Qualifications Wales (2020e), SQA (2020b), Welsh Government (2020d), WJEC (2020a,b)

The policy began to unravel in August following the publication of moderated National 5, Higher and Advanced Higher results in Scotland on 4th August. Immediately, the media focused on the 'downgrading' of learners' outcomes between the grades estimated by centres and the final standardised grade (Matchett 2020). Although, as shown in *Table 1*, equality and fairness considerations formed part of the principles intended to underpin the awarding of grades in 2020 – and SQA publishing statistics that showed that attainment gaps had narrowed compared with previous years (SQA, 2020d) – the higher proportion of grades adjusted amongst students from more deprived areas caused discontent. This was reinforced by personal stories of downgrading, including some from learners who had achieved better grades in 'prelim exams' taken prior to school closures, and those who claimed to have missed out because the model did not allow for the possibility of learners achieving a grade higher than those achieved at the centre in previous years (BBC, 2020a).

The damage to the credibility of the moderation process proved terminal. On 11th August, John Swinney, the Scottish Education Secretary, announced that all learners would receive the better of their moderated grade and the estimated grade provided by their centre, with provision made to resource the resulting increase in demand for university and college places (Scottish Government, 2020a). Following that announcement, stakeholders began to make links to the situation in the other nations (Flockhart, 2020), and similar issues emerged when England, Wales, and Northern Ireland published their AS and A-level results on 13th August. Governments all changed their policies within a week, prior to GCSE results day. In England (Ofqual, 2020e) and Northern Ireland (CCEA, 2020d), learners were awarded the higher of their CAG or their standardised grade for all general qualifications. In Wales, the same was applied as a second stage of policy change, but in a first stage of policy change following the publication of A-level results, the Education Minister had already given a direction for A-level candidates to receive no lower than the AS grade they had achieved in a previous examination series (Qualifications Wales, 2020f).

The aftermath: implications for future assessments

Once grades had been re-processed and published, the post-mortem began almost immediately. Different approaches to the scrutiny of policy decisions, and to the next summer examination series in 2021, were taken across the UK. In England, the Government resisted calls for an independent review, but the House of Commons' Education Committee published their views in November 2020, raising concerns regarding the relationship between the Department for Education and Ofqual, and a perceived lack of transparency and scrutiny in the development of the algorithm used to standardised grades. They also stated their view that exams must go ahead in 2021, with a mechanism in place to account for differential teaching and learning loss caused by local 'lockdowns' (House of Commons Education Committee, 2020). The Information Commissioner's Office (2020) and the Office for Statistical Regulation (2020) also undertook inquiries focused on the use and processing of data in standardisation and grading decisions. While the former did not publish findings, the latter concluded that regulators and exam boards had acted with integrity, but that ensuring greater openness and transparency, as well as maintaining rigour and quality assurance, may have increased public confidence in the resulting grades.

At the end of 2020, a full examination series was due to take place in England in summer 2021, with examinations – considered by the Government to be the "fairest form of assessment" – adapted to account for the loss of teaching and learning time and to reduce assessment burden for examination centres (schools and colleges) and learners (Department for Education, 2020c, p.1). Public examinations were also to take place in Northern Ireland, where the Education Minister emphasised the reduction of assessment burden caused by amendments to 2021 assessments compared with normal series (Northern Ireland Department for Education, 2020). In both England and Northern Ireland, the policy focus was therefore on maintaining, as much as possible, usual arrangements for awarding qualifications, but examinations were ultimately cancelled due to the

emergence of a second Covid-19 wave in the winter of 2020/21 (Department for Education & Ofqual, 2021; BBC, 2021).

Different decisions were taken in Scotland and Wales. A rapid review commissioned by the Scottish Government (Priestley et al., 2020) recommended that National 5 examinations should be replaced with centre estimated grades, based on assessments validated by SQA. The recommendation for Higher and Advanced Higher examinations, aimed primarily at 17- and 18-year-olds, was that assessment should go ahead, with modifications made to maximise teaching and learning time. The review also recommended that a commitment be made to embedding equalities and children's rights into all aspects of the qualifications system, to ensure just decisions for each individual student. These recommendations were all accepted (Scottish Government, 2020b). In December 2020, Highers and Advanced Highers were also cancelled due to continuing disruption to teaching and learning (Scottish Government, 2020c). An OECD (2021) report led to the Scottish Government committing to eventually replace SQA (Scottish Government, 2021).

The Welsh Minister for Education commissioned an Independent Review panel to review the decisions made through the 2020 exam series and make recommendations for 2021 focused on learners and their progression as well as the need to maintain standards and the integrity of qualifications (Welsh Government, 2020b). Their interim report (Casella et al., 2020) focused on arrangements for the 2020-21 academic year. In contrast to the decision made in England, the panel concluded that any examination series would produce results that reflected inequities in access to teaching and learning opportunities. This led to a Ministerial decision to replace examinations for GCSEs, AS and A-levels in 2021 with a system of "teacher-managed assessments" (Welsh Government, 2020c). The panel rejected concerns about equity and comparability arising from the approach and stated that 'fairness' to learners being assessed in 2021 must be prioritised above all other considerations. In both Scotland and Wales, recommendations were also made for examinations beyond 2021.

In all four nations, the events of 2020 are likely to leave an enduring mark on how the population of the UK, and the teaching and assessment communities more specifically, view educational assessment. The approaches developed to award grades can provide insight into how we understand the key principles of assessment and how they interact. They also reflect the effect of divergence in qualification design, with each UK nation's government beginning to take different decisions on their qualification systems.

Research aims and questions

The aim of the study is to explore the short- and potential long-term effects of policies relating to the summer 2020 general qualification examination series in Wales, with respect to the key assessment principles of validity, reliability, fairness and comparability. Wales is an interesting case study due to its similarities with other UK nations in differing aspects of education systems: England for levels and brands of qualifications, Northern Ireland for the modular A-level qualification structure, and Scotland for curriculum principles (Donaldson, 2015). Thus, all four nations are usual reference points throughout this study, reflecting the fact that the four systems represent 'variations upon common themes' as well as the interdependence between the systems (itself illustrated above) (Raffe et al., 1999). In this sense, Wales is the main case study, while drawing on other UK nations for further elucidation and comparison, adding further insight in responding to the research questions (Yin, 2018).

Three research questions are considered. In Section 3, I consider how the key principles of assessment were balanced differently in the standardisation models used in 2020, compared with normal examination series, and well as the alternative approach based on awarding CAGs argued for by some stakeholders. Decisions about how these principles should be prioritised are an integral element of designing an assessment system, and the need to develop a new system between March and results publication in August 2020 meant that these considerations were made much more quickly than would normally be the case. Within this section, I evaluate critically the validity

arguments and the relationships between different forms of fairness and comparability in each approach – particularly given the ways in which conceptions of fairness were so often evoked in public discourse in 2020 and thereafter. Consideration is also given to inferences made within the three approaches, the definition of standards used, and the strengths and vulnerabilities of each.

Given the focus on CAGs and alleged 'downgrading' in summer 2020, in Section 4 I review the grades submitted by schools and colleges. I discuss the usefulness of these grades in each of the approaches set out in Section 3, evaluating how plausible the CAGs were and how that affects the balance of key assessment principles in each approach. Considering this analysis, I then review the extent to which it was possible to produce a set of grades that met all stakeholders' expectations at the examinee/student and system/cohort levels.

Finally, within my conclusions, I examine how the debates arising from 2020 could affect future developments in UK general qualifications. With reference to the Priestley review in Scotland and the Independent Review in Wales, I draw on the literature on paradigms of assessment (Baird, 2018a) to examine whether proposals for future assessments might require a further reconsideration of validity, reliability, fairness, and comparability.

Section 2: Research methods, ethics and positionality

Conducting research as an 'insider'

This research is a critical evaluation of the positions taken during the summer 2020 examination series, drawing on academic work from the field of education and assessment, and the broader public policy literature. I have taken a critical realist perspective, drawing on the social realism tradition within the sociology of education. This perspective posits that *intransitive* social objects – such as assessment models – can be invented into reality through underlying social processes (Young, 2008). People can create knowledge about such objects, but this knowledge is *transitive*, reflecting the conditions under which it is produced (Moore, 2013). This research represents my attempt to understand the assessment models that have been proposed and implemented since the start of the Covid-19 pandemic.

Awareness of my position and background was a crucial element of undertaking this research. I completed the study as an 'insider' with direct involvement in the design of the standardisation processes that were put in place in England and Wales in 2020, due to my responsibilities for standards at the WJEC examination board. I was involved in discussions with stakeholders throughout that process, including the subsequent reviews, discussions and developments which followed. The role of the insider-researcher is well-documented (*cf*. Costley et al., 2010). Gray (2018) applied this body of knowledge to conducting research within exam boards, noting that being reflexive and reflective can bring authenticity to research and bulwark it (and the researcher) against charges of a lack of objectivity. Reflexivity is not something to 'be' but is an essential element of conducting social research, demanding a recognition of the fact that we help to construct the social world as we engage with it (Atkinson, 2015). Reflective practice requires us to recognise our reactivity, and given that we cannot eliminate it, to be aware of how it shapes the development of our research.

Gray (2018) refers to Stenhouse's (1975) concept of the 'extended professional' engaged in autonomous self-reflection. My primary purpose in choosing the focus of my research was to extend my understanding of an ongoing debate which not only continues to impact on the lives of hundreds of thousands of young people, but one in which my work – and that of my employer – has come under considerable criticism. By changing my role from insider to insider-researcher, I aimed to become less reactive and more analytical when considering the options available for awarding grades, gaining a more critical understanding of my own perspective and those within my organisation.

My position and theoretical perspective shaped the choice of research methods. The primary requirement in the data collection phase was to collect sufficient information to describe proposed assessment models and identify key critiques. I chose not to gather primary qualitative data via fieldwork – interviews, observation or focus groups – from those involved either in standardisation or the wider discussions around relating to my research topic. Given that the research was started while decisions on the 2021 assessment model for Wales and England were still being agreed, engaging key stakeholders directly in the research process could have created unnecessary reputational risks for my organisation. I chose not to use my 'insider knowledge' of others' views shared with me and others in my organisation over the course of the 2020 grading process, partly for the same reason, and partly to ensure that the findings of my research are more easily verified.

Document analysis

Given the focus of my research, the documents required were those that set out relevant public statements from key stakeholders involved in those national public debates on assessment. These were gathered from relevant organisations between 1st March 2020 and 30th June 2021, using the sources listed in *Table 2*. The website news sections and Twitter feeds of Welsh Government, Qualifications Wales, WJEC and all major school and college teachers and leaders' unions, opposition political parties, and relevant expert bodies were reviewed. Transcripts of Senedd

(Welsh Parliament) sessions were included, along with reports and publications from each of the organisations listed. All published statements related to assessment in the 2020 examination series and thereafter were captured.

Searches were also conducted on the websites of Wales' major news outlets using the search terms 'examinations', 'exams', and 'grades' to gather wider policy-related news and opinion pieces; articles were then reviewed to ensure they related to assessment during the pandemic. The capture of views of leaders, teachers and students was limited to those sources; a deliberate choice to make the research more manageable and to focus the analysis primarily on stakeholder debates.

Table 2: List of sources for document analysis

Organisation type	Organisations	Website	Number of sources
Government	Welsh Government	gov.wales/announcements	23
Senedd	Senedd	record.assembly.wales	16
Regulator	Qualifications Wales	qualificationswales.org/english/news/	8
Exam board	WJEC	www.wjec.co.uk/articles/	31
Opposition political parties	Plaid Cymru Welsh Conservatives	www.partyof.wales/updates www.conservatives.wales/news	12 5
School leaders' unions	NAHT Cymru ASCL Cymru	www.naht.org.uk/About-Us/NAHT-Cymru www.ascl.org.uk/News/Our-newsand-press-releases	3 10
Teachers' unions	NEU Cymru NASUWT UCAC	neu.org.uk/neu-cymru www.nasuwt.org.uk/advice/wales.html www.ucac.cymru	1 1 4
Media	Wales Online Daily Post Nation Cymru Golwg BBC	www.walesonline.co.uk www.dailypost.co.uk nation.cymru golwg.360.cymru www.bbc.co.uk	6 7 14 3 1

This approach is unobtrusive and non-reactive (Webb et al., 1966), as the materials are not shaped by subsequent thinking and should therefore reflect the writers' perspectives at the time. This is a strength in research settings where the focus of interest is changing quickly (Bryman, 2012). In Section 3, key principles of assessment are set out and used as the theoretical frame for analysing the documents. The analysis was used to produce a *description* of a series of validity arguments for

approached developed and used in 2020. *Critiques* of each model were also sourced from the documents, and the academic literature, to compare and critically evaluate each validity argument.

Research quality criteria

A critical realist perspective requires careful consideration of quality at all stages of the research process. Scott (1990) identified four criteria for assessing documentary evidence: authenticity, credibility, representativeness and meaning. By limiting the selection of documents to public statements from key stakeholders, questions of *authenticity* and *credibility* relating to the supplier of the text itself are largely resolved. The accuracy and sincerity of the texts were considered within the analysis, with careful reflection on my position within that process. Maxwell (2017) stated that the validity of research is determined by the extent to which the inferences made in the research are supported by the design, methods, data and analysis; due attention was therefore paid in the analysis phase to prevent misinterpretation.

Representativeness is considered for both proposition and critique, to establish the prevalence of each viewpoint (Martin, 2018). This ensures that any divergence of views within and between each model is drawn out. *Meaning* relates to intention; any content related to critique was assumed to be intended to persuade. The forms of documentary evidence used in this research are not inert and may have been produced for the purposes of persuasion in a contested policy space, meaning that documents can become active agents in their own right (Prior, 2016). For this research, however, documents are treated as a resource, as the focus is on their content and meaning, rather than a topic in themselves (Prior, 2008). The political nature of discussion about qualification grades meant that the analysis of the documents must focus not just on the substance of critique, but also – with reference to the theoretical framework set out in Section 3 – consider the values and principles that underpin those statements (Gray, 2018).

Analysis of grade data

I also undertook an analysis of candidate grade and mark data, to enable a critical examination of one of the key areas of differences between the models proposed in 2020: the plausibility assigned to CAGs submitted for use in standardisation processes. Limited documentary evidence was available to evaluate this point. The research is therefore a mixed-methods study in the sense that I combine qualitative data collection and analysis, and quantitative analysis of secondary data, to increase the breadth and depth of my understanding (Johnson et al., 2007).

Data from students in Wales entered for Qualifications Wales-regulated A-levels was used for this element of the research, for three reasons. First, it relates to my study's main nation of interest. Secondly, the structure of the qualifications lends itself to such an analysis: most learners in receipt of an A-level CAG in Wales completed AS units in 2019 which contribute to the A-level, and most centres had learners entered for the same qualifications in previous examination series. This meant that those setting the CAGs had reliable assessment evidence available to them to inform that process. This was not the case for other qualifications, meaning that centres engaged in this process had an advantage. The third reason was access: all Wales A-levels are awarded by WJEC, and as a WJEC employee, I was able to access the data via the internal approval policy for utilising the organisation's data for research purposes. More details of the purpose of this analysis, the specific methods of data collection and analysis used, and the extent to which these findings can be said to be generalisable, are set out in Section 4.

Ethical approval

Ethical approval for this research was requested from the University of Oxford School of Education's Departmental Research Ethics Committee (DREC) on 2nd March 2021, and ethical approval was given on 8th March 2021. A copy of the application is provided in Appendix 1, including notice of the approval given by WJEC to undertake the project and to use the candidate-level data needed for the

analysis undertaken in Section 4. The British Educational Research Association *Ethical Guidelines for Educational Research* (2018) were applied at each stage of the research process.

Section 3: Key principles of assessment in awarding grades in 2020

The key principles of assessment

Definitions of the key principles of assessment form the basis of my evaluation of the validity arguments. Of all the principles, *validity* is considered by many to be the most important concept in educational and psychological assessment (Newton & Baird, 2016). The pre-eminent stance within educational assessment is that validity is a property of test score meaning, interpretation and use, as well as the resulting consequences (Messick, 1995). The Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014, p.11) define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests".

Other key principles of assessment link with validity. Often evoked in debates about educational assessment, *fairness* has multiple meanings. Gipps and Murphy (1994) distinguish between the fairness of assessment practice and the interpretation of group differences in achievement; if the assessment process introduces unfairness, it is not valid to interpret any difference in grade outcomes between groups as the result of differences in levels of attainment between those groups. It is important to differentiate *equality of opportunity*, which requires only that no unfairness is introduced via assessment; and equality of outcomes between groups, referred to hereafter as *equity*.

Nisbet and Shaw (2020) identified several 'senses' of fairness. The *formal* sense means correct application of a rule, without any assessment of the appropriateness of the rule. The *implied contractual* sense relates to the legitimate expectations of those affected, arising from awareness and application of regular practices. The *relational* sense means treating like cases alike. The *retributive* sense means that an outcome is fair if it is justified. Nisbet and Shaw point out that this sense can be utilised in a value-neutral way (relating to the demonstration of attainment) or an evaluative way (relating to deservedness arising from effort). Finally, *consequential* (concern that

the outcome of an assessment could lead to unfair future decisions) and *retrospective* (outcomes resulting from past unfair decisions) senses add a temporal element to the concept.

The implied contractual and relational senses link fairness with *comparability*. Comparability has a broad meaning, encompassing many forms of comparison between elements of assessments (Elliott, 2011). The primary focus in this study is at qualification level, focused on the extent to which grades awarded in 2020 and previously can be considered transferable in relation to their primary uses, and the extent to which that matters. The types of comparability of particular interest for this study are therefore inter-subject comparability, inter-jurisdiction comparability (between the four UK nations), comparability between centres, and comparability of the meaning of grades over time.

Reliability is the extent to which variation in a score (or outcome) is the result of differences between individuals in their respective levels of the underlying attribute being measured, rather than due to measurement error (Furr, 2018). Reliability is therefore considered to be a necessary but insufficient element of validity (Kane, 2013): the less we can rely on an assessment outcome as reflecting the attribute of interest, the less likely it is that we should use that outcome as intended (Winkley & Cresswell, 2012). Which forms of reliability we choose to focus on reflects the nature of the assessment and our perspective on it (Baird, 2018a).

Authenticity relates to two related concepts of how we interpret assessment performances; first, the assertion that the performance belongs to the student ('authorship'); second, how the performance was generated ('assessment conditions'). This information facilitates inference of a student's level of attainment in the domain of interest[1]. For the latter, authenticity is a property

[1] This is distinct from another use of the term 'authentic assessment', used to refer to changes in the nature and methods of assessments, away from written texts of knowledge and towards contextualised and meaningful assessment of learning in real-world contexts (Pellegrino et al., 2001). This definition is not relevant to this study.

attached to changes or differences in the quality of performances, and therefore links closely with comparability. Performance gain is authentic if teaching of content improves ('realignment'), but inauthentic if the performance gain occurs because teachers have learned how to prepare students for assessment structures or formats ('adeptness'), provide strategies for scoring marks that their learning doesn't deserve ('coaching') or focus their teaching on what is most often assessed ('reallocation') (Newton 2020).

The *manageability* of an assessment process can impact on all the key principles. Any assessment procedure creates burdens – on assessment designers, learners, examiners, awarders and users – and ensuring that there are sufficient resources to deliver an assessment is an important factor in meeting the key principles. The purpose of general qualification grades in 2020 was intended to be the same as in previous and future years: to allow progression to further study and employment (Department for Education, 2020b; Welsh Government, 2020d). This means it is legitimate to evaluate various approaches to awarding grades on an equivalent basis, through the lens of the key principles set out above. In the next section, I set out my framework for doing so.

A framework for comparing validity arguments

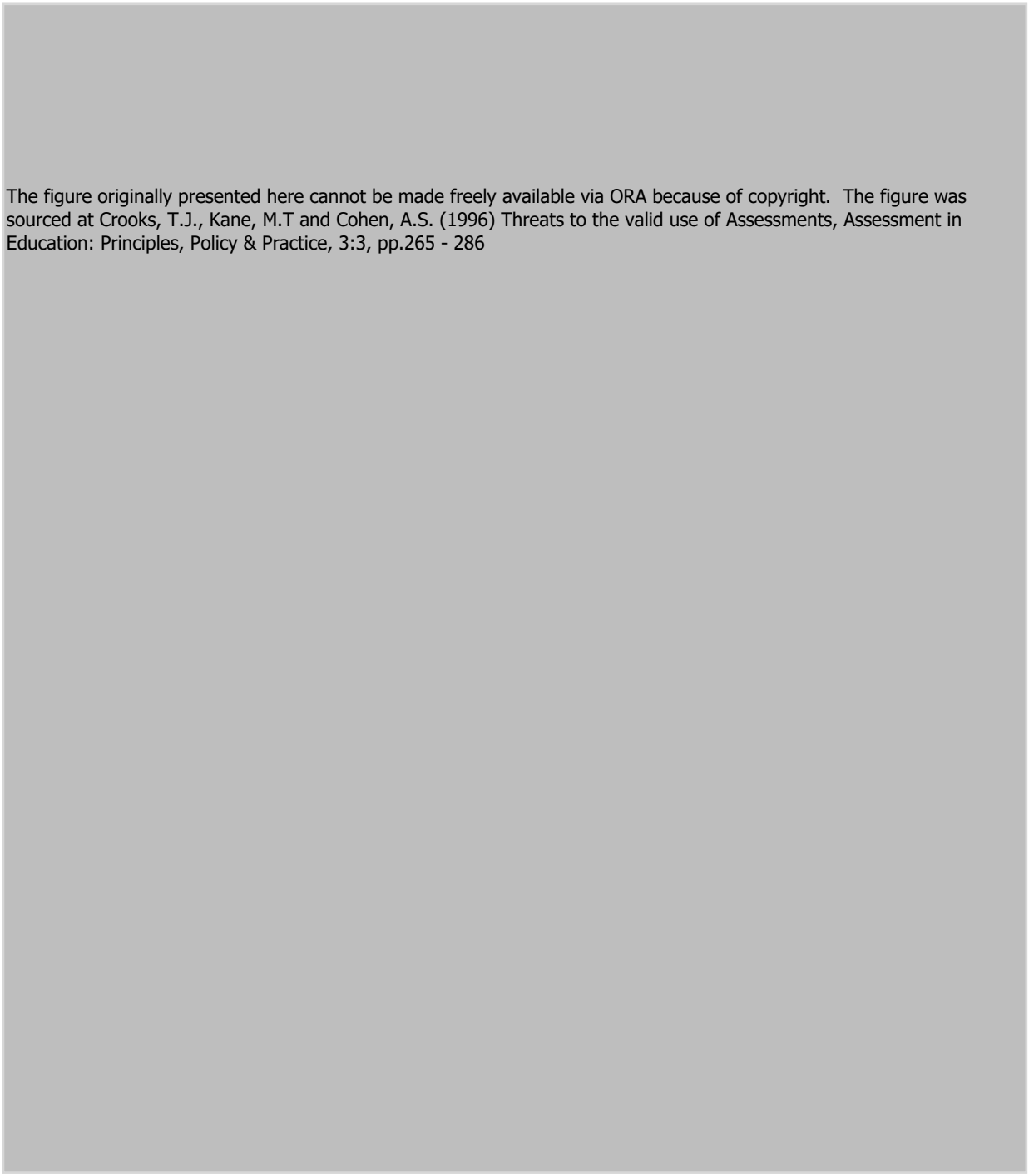
Various frameworks and approaches exist for evaluating validity. Kane's (2013) interpretation/use argument fits with the Standards for Educational and Psychological Testing definition of validity by proposing that each interpretation/use should be validated, highlighting the need to consider the inferences made between elements of the assessment process when evaluating validity. The approach was heavily influenced by Toulmin's (1958) model of inference: general rules called *warrants* are used to infer a *claim* from a particular piece of *data*. *Backing* is the evidence used to support the warrant.

Newton (2013) argued that, the selection of a framework is less a question of what it does as how it facilitates investigation in the given context. For my study, a stage-based approach enables a

breakdown of the various approaches to awarding grades into their respective parts. The stages are grouped into four phases, which are discussed in more detail in the following sections and set out in *Table 3*: performance generation, construct representation, awarding, and consequences. The first five inferential stages cover the 'performance generation' and 'construct representation' phases and match the first five stages from Crooks, Kane and Cohen's (1996) validity chain model, which depicts assessments as a chain of 'links'. An assessment is said to be only as strong as the weakest link in that chain. In line with Crisp and Shaw (2012), a single warrant is identified for each stage alongside the relevant backing. To aid comparison, the assessment process used for general qualifications in Wales in pre-pandemic years (and elsewhere around the UK) is set out, alongside models for 2020.

The third phase – awarding – is a key inference within UK general qualifications (Newton, 2011) and relates to evaluation link of the validity chain model. The final phase – consequences – deals two stages: progression, and the broader impact of the grade. This integrates elements of Newton and Shaw's (2014) 'neo-Messickian framework' for evaluating testing policy, incorporating both the primary and secondary impacts of assessment, focusing on the interpretation and use of assessment outcomes, and wider societal considerations.

Table 3: Inferential stages for general qualifications (based on Crooks et al., 1996)



Performance generation

Administration

The performance generation phase consists of the administration and scoring links of the validity chain model. *Administration* covers the completion of the task by the candidate, and the context in which it was completed. In a normal examination series, rules for assessment administration are set by the exam boards, to ensure consistent and secure delivery and thus ensure the integrity of the process (JCQ, 2020a). These rules are designed to ensure that, except for those students requiring 'reasonable adjustments' under the Equality Act to prevent disadvantage in accessing assessments, students receive the same assessment experience (JCQ, 2020b): both the formal and implied contractual senses of fairness described previously are invoked by their consistent application. The rules cover security of assessment material, accommodation and invigilation, providing access arrangements for those with specific requirements, and malpractice. Specific rules may also be set out in the qualification specification.

The administration stage also covers other factors that could affect the conditions in which performances are gathered, such as question paper errors. Student motivation is another factor (Crooks, Kane & Cohen, 1996): in a normal examination series, learners should know that what they produce in an assessment will contribute to their qualification outcome. Again, this links to the implied contractual sense of fairness.

Marking

Once performances are gathered, they must be scored, or *marked*. In a normal examination series, student performances collected in examinations are marked online and anonymously by a group of subject experts (usually teachers). To do this, they are trained in the use of an agreed 'mark scheme', which sets out how to credit item responses. Quality controls are imposed through the marking period: where marking is outside a tolerance relative to the marking of a senior examiner,

the examiner is retrained or stopped from marking (AQA, 2021). Inter-rater reliability is therefore a key concern (Baird, 2018a), and Ofqual have recently begun to publish statistics on the effect of marking reliability on classification accuracy (an estimate of the proportion of learners who receive the 'correct grade', based on a statistical analysis of marking data (Rhead, Black & Pinot de Moira, 2018)). This has led to public criticisms that the value of exams is undermined by the Ofqual estimate that one in four grades is not 'correct' (Sherwood, 2020). Mark schemes are used to standardise what features of performance are credited. Mark scheme writers must reflect the subject domain of interest by balancing holistic features, such as creativity and richness, and an analytic approach focusing on specific features (Crooks, Kane & Cohen, 1996).

Some components of general qualifications, such as the production of art pieces, are internally-assessed: marked by teachers in centres, using exam board mark schemes. A sample of student performances and marks are then being moderated by exam board appointees, and marks adjusted if the marking standard is not appropriately applied (Pearson, 2021). Once marking is complete, a student's mark may be adjusted to account for issues which have affected their performance, such as missing an assessment, minor illness or bereavement. This process is known as special considerations (JCQ, 2020c). The formal sense of fairness (as rules are in place to ensure consistent application), and the retributive sense of fairness (in that candidates would otherwise have received a lower mark than they would otherwise have achieved) are pertinent here.

Performance generation in 2020

For unitised qualifications, such as A-levels, most students already had marks and grades banked which would have contributed to outcomes if exams had taken place. For linear qualifications, such as GCSE English Language, some non-examination assessments covering aspects of the construct may have been part-completed but not subjected to external marking or moderation. Otherwise, given the circumstances in which centres were forced to close in March 2020, it was inevitable that

CAGs would be based on a variety of sources of evidence already available to centres, each of which is likely to have been administered in slightly different ways.

JCQ (2020d) provided guidance on the level of confidence that should be given to difference sources of evidence, emphasising that centre-supervised assessments should be given the greatest weight in making the holistic grading judgement. In place of the usual backing for common administration practices, guidance on accounting for the assessment conditions was also provided *(cf*. WJEC, 2020c), thereby covering both concepts of authenticity. Mark schemes were provided to guide the marking of assessments from previous series that had been utilised by centres, supporting marking reliability. Unlike a normal series, however, this marking was not standardised by exam board examiners. Beyond this, however, there were no differences between standardisation and a CAGs-only approach in this phase. Both arguments rely on the same evidence.

Construct representation

Aggregation

The construct representation phase consists of three stages. Combined, these stages have the purpose of obtaining a final mark for each candidate, that represents their level of performance in assessments that, in turn, represents the subject domain of interest.

Aggregation occurs within and then across assessment components, to reach a final qualification mark. For normal examination series, aggregation rules are determined in qualification specifications, to ensure consistency (and therefore fairness in the formal sense) for all candidates. Consideration of aggregation is particularly important where the subject domain of interest is composite rather than unidimensional, and therefore the relative weighting of each attribute is a subjective decision (Maul, 2013). In subjects that require assessment in both practical skills and knowledge, such as music or physical education, the intended weighting of each skill within the

aggregated mark may not reflect the actual weighting if those with stronger practical skills are attracted to that subject.

Generalisation and extrapolation

Kane (2013) links *generalisability* to reliability: a check that the information gathered is sufficient in depth (number of tasks) and consistency across learners that we can use it as a common basis for making judgements. In a normal examination series, each component contains several tasks (items) that are combined, for all qualifications that have more than one component. The next stage, *extrapolation*, is a check that the information gathered is sufficient in breadth, that the subject domain of interest is covered sufficiently, and content sampled appropriately, so that the meaning of candidate marks can be inferred as representing the domain (Crooks et al., 1996). Generalisability and extrapolation relate to validity in the narrower sense of the term – that we are measuring what we think we are (Borsboom et al., 2009) – as well as the legitimate expectations sense of fairness, that all learners are being assessed against the same domain of interest, in this series and between series.

Construct representation in 2020

Although aggregation could not take place as normal in the 2020 grading process, WJEC (2020c) provided guidance to reinforce the usual backing for the warrant at each stage, advising that teachers should ensure that the assessment objective weightings set out in each qualification specification were replicated when making grading judgements. To resolve issues with generalisability, the same guidance recommended forms of evidence teachers should use to support judgements against each assessment objective; it was accepted that centres would have different sources of evidence to weigh up. JCQ (2020d) set out what actions were required where insufficient evidence was available for a student, including potentially not awarding a grade. As with the performance generation phase, differences between standardisation and a CAGs-only approach

in this phase were limited, as the same activities underpinned both, but the usual backing for a validity argument was not present.

Awarding

By the end of the construct representation phase, we usually have a mark for each learner that reflects the quality of their performance, on a common scale, and have confirmed that the scores appropriately represent the 'target domain'. The next phase – awarding – requires a judgement about the meaning of scores. In UK general qualifications, statistical evidence is reviewed alongside examiner judgements on samples of learner performances to set grade boundaries that maintain standards between examination series. The grade boundaries are then applied to the whole cohort entered for the examination (Taylor & Opposs, 2018), on the assumption that the mark distribution fairly and reliably reflects the relative level of attainment achieved by learners.

The conceptualisation of standards within the awarding process is much-discussed, but definitions abound, often resulting in those involved in debates in standards talking at cross-purposes (Baird & Gray, 2016). For my framework, it is necessary to clarify two issues: what form of standard is being maintained (and not being maintained), and by what method or definition standards can be said to have been maintained.

<u>What standard is being maintained?</u>

Newton (2011) stated that an examination standard is "[x]-referenced", in the sense that the standard is defined in terms of whatever is the [x], which is also the focus for comparability. Several types of standard referring to the *features* of the assessment process can be found in the literature (Opposs & Gorgen, 2018): content standards refer to the programme of study; demand standards to the level of demand of the assessment tasks; marking standards to the level of credit given for a response (as the same response can be given many marks or few marks depending on the structure and level of expectations set within a marking rubric). Confusion begins when we get to the

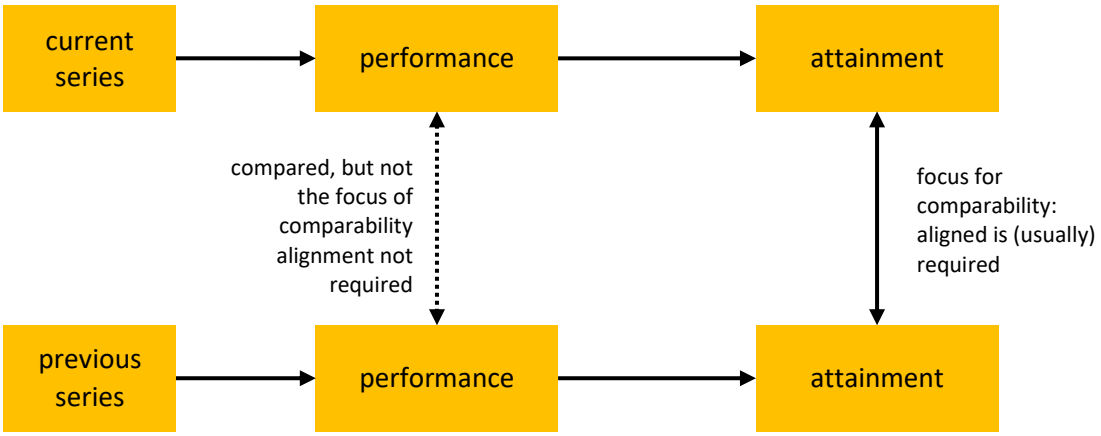
awarding stage, however: terms such as examination standards, assessment standards, performance standards and grade standards (Baird & Grey, 2016; Opposs & Gorgen, 2018) are used interchangeably seemingly without consideration of whether they mean the same thing.

The most important distinction is that between performance and attainment. Coe (2010) critiques a strict interpretation and comparison of performance as a basis for maintaining standards (and awarding grades) without either accounting for the context in which the performances were produced. In practice, as Coe says, a comparable performance approach would require a consideration of the relative demand of the assessment but may not consider wider issues. Newton (2010a) took this further, stating that, while the judgements made at award were about comparing observable performance, the purpose of the award was to link *attainment* constructs from series to series. As a result, Newton (2011) proposed that the UK examination system is based upon 'attainment-referencing': seeking to align the meaning of grades between series as representing a level of skills, knowledge, and understanding of the subject curriculum.

This is illustrated in *Figure 3*. A committee of senior examiners conducts the award, a nod to the *conferred power* definition of standard-setting that posits that using subject experts can buttress the process against critiques of the validity of the inferences made (Baird, 2018b). When awarders are making judgements about candidate performances on a particular mark in a normal examination series, awarders will be given 'archive scripts' for reference – candidate work on the grade boundary from a previous series. Just as awarders in a previous examination series will have inferred that these archived performances represented a level of attainment, so awarders will be asked to use the presented performances to make a similar inference about a given mark in the current series. While performances between series are compared, therefore, the question that awarders are asked is not to compare performance, but to establish if the performance is worthy of the higher grade being considered (Taylor & Opposs, 2018). Answering this question requires

performance to be *contextualised* within the current series to infer a level of attainment from it, including consideration of the assessment conditions in the previous series.

Figure 3: Comparison and comparability between performance and attainment standards: a conceptual map



The focus for comparability is the attainment standard – this is the warrant for awarding. Within a series, the formal and relational senses of fairness are managed by applying grade boundaries to a mark distribution, in which any specific issues at learner level are dealt with by adjusting marks prior to the awarding stage. Aligning attainment ensures formal and relational fairness in terms of consistently maintaining the meaning of grades across nations and series. The retributive sense is used in a 'value-neutral' way: attainment, not effort, determines the grade. The implied contractual sense of fairness is met by retaining a similar process across all awarding bodies across all subjects, years, and nations. The awarding process is therefore key to maintaining validity across the assessment process, by dealing with fairness and comparability issues in a consistent manner, thus providing the backing for the awarding phase of the validity argument.

Making these inferences is a complex task. We can distinguish here between three sources of complication when considering the delivery of an assessment. First, the *assessment conditions* – task familiarity, repeated administration and reliability, other aspects of task administration and control. Second, the *assessment task* – changes in the content sampled, task demand, or marking

standard. Lastly, sources related to the wider *societal context*, which are less common, but which include assessment during a pandemic. All three relate to the performance generation phase of my framework.

I use the term *performance standard* to refer to the features of candidates' assessment evidence that demonstrate a particular level of skills, knowledge and understanding. *Attainment standard* refers to the level of underlying skills, knowledge and understanding obtained by the learner. In a normal examination series, the warrant for the awarding stage is that the attainment standard is maintained, so that the grades awarded represent an equivalent level of attainment between learners in the same series, and from series to series. This in turn enables validity in the subsequent stages. I therefore use the term attainment-referencing to reflect the fact that the standard is defined in terms of attainment, and not in terms of the methods used to maintain that standard. This is explored further in the next section.

<u>What method is being used?</u>

Newton's argument that attainment-referencing ought to be the aim of an educational assessment awarding process is itself underpinned by a belief that performance alone is not sufficient to award a grade. As a field, psychometrics is built on the belief that a construct or trait of interest exists, underlies performance, but is latent (i.e., not directly measurable) (Furr, 2018), even if the level of the construct can be increased through teaching and learning. In other fields, such as vocational qualifications, observation of performance is considered sufficient as a basis to ensure competence (Wolf, 2001). Baird (2018a) characterises these positions as paradigms of assessment. The third paradigm – curriculum-based assessment – is where UK general qualifications sit. Construct validity – how the assessment relates to the curriculum – and inter-rater reliability are of primary importance.

The information that we require to make grading judgements is therefore shaped by the position we take on what constitutes evidence of achievement, and what we need to satisfy in making that argument. Different methods have been developed for maintaining the attainment standard. Baird's (2018b) ecological system of standards considers methods at different levels. *Examinee-level* approaches focus on the meaning of collated performances produced by each learner. Examination system approaches take the cohort of assessment takers as the basis for standard-setting – positioning the learners in relation to each other and setting cut scores accordingly. *Social and cultural context* approaches take the examination system definitions of standards a stage further, by accounting for examinee characteristics that can affect performance. Finally, *systemic* approaches attempt to account for all the previous definitions, to balance political and technical risks, accounting for the cohort entered for the qualification and the context in which they performed.

Baird positions attainment-referencing as a systemic definition – but while Newton's (2011) exposition of 'attainment-referencing' reflects that paper's focus on the methods used in the awarding of A-levels, its definition of the term does not specify methods for maintaining standards – only that consideration of performance alone cannot be sufficient as a basis for awarding grades that reflect attainment. This illustrates the known challenges around the reliability of absolute judgements of candidate work on grade-worthiness, and in particular an optimism bias (Ofqual, 2020f).

Within a systemic approach, other evidence can bolster the reliability of awarding decisions and cover other definitions of standards that may otherwise create risks. In general qualifications awarding, statistical predictions of outcomes for the largest age group amongst examinees, based on their prior attainment, are used to identify grade boundaries that would lead to 'comparable outcomes' (Cresswell, 2003) for that group. These boundaries are used as the starting point for awarder judgements; the prediction is also used to review outcomes at the end of the awarding

process (Ofqual, 2019c; Qualifications Wales, 2019a). This is a causal definition of standards – that comparable standards occur when equivalent grade boundaries are the same in terms of the causes of their attainment (Newton, 2010b) – which sits under the social and cultural context level within Baird's framework. In this instance, prior attainment is treated as the predominant cause.

At AS and A-level, the prior attainment measure is commonly calculated across England, Wales, and Northern Ireland, illustrating the close link between the causal definition and Coe's (2010) concept of statistical comparability. This secures inter-jurisdictional comparability in qualifications that learners use to compete for university places. A systemic level approach to maintaining attainment standards is therefore operationalised using a mixture of performance judgements at the examinee level and prior attainment-based statistics referring to the primary cohort, at the social and cultural context level.

The concept of 'comparable outcomes' was initially conceived to prevent a fall in outcomes in the first examination series after the introduction of a reformed qualification specification: a 'sawtooth effect' that sees general qualification performance improve as test familiarity (part of the assessment conditions) increases (Cuff, Meadows & Black, 2019). This led Newton (2020) to distinguish between a 'comparable outcomes' period in which the meaning of grades is temporarily sacrificed to ensure fairness in consequence, to account in a genuine (but assumed temporary) reduction in both performance and attainment caused by a change in curriculum and/or assessment, and a period thereafter in which grades can be interpreted consistently. Put another way, the consequential and retrospective senses of fairness are protected, at the expense of the relational sense as it pertains to attainment – cases with similar levels of attainment in different years are not treated the same – and therefore the attainment standard itself. The retributive sense changes too: attainment is not truly aligned, giving away to an evaluative conception: deservedness. Statistics support this approach, but the role of awarder judgements (and therefore candidate

performance in awarding) changes, requiring them to consider the context in which performances were generated (Ofqual, 2019c).

Awarding in 2020: accounting for authenticity

In a normal series, the awarding warrant is based on a maintenance of the attainment standard, backed by procedures, reviewing of assessment performances and information gathered on the context in which those performances were gathered. The content standard (and coverage of specification content) is also assumed to be equivalent across all examinees. As all schools were closed in March, much of the school year had already passed, and therefore content coverage was also assumed to be sufficiently comparable across all schools for content standards to be unaffected in 2020. In the absence of an awarding committee, conferred power for awarding grades – the inference from performance to attainment was given to teachers, who provided the CAGs (and rank order positions) representing the level of attainment they believed the student would most likely have demonstrated if exams had taken place. While teacher assessment can refer to any assessment process in which the teacher is responsible for the inferences that in general qualifications are made at the awarding stage (Harlen, 2004), the production of CAGs was not the same as 'teacher assessment' as usually described in the literature, as a deliberate and planned practice (Johnson, 2013). The 2020 process required a prediction based on assessment evidence that had originally been captured for a different purpose; as stated above, the performance generation and construct representation phases were not common across all students. The single scale of marks that would exist in awarding in a normal series, in which relative performance in different aspects of the domain would be weighted appropriately and consistently at the aggregation stage, was not present. At best, students were on a scale (of sorts) within, and determined by, their own centres.

As noted previously, guidance stated that authenticity concerns - both authorship and assessment conditions – were to be accounted for when making the grading inference. Grade distributions at

the centre level differ between years (Qualifications Wales, 2019b), reflecting genuine, authentic differences in attainment levels, and depending on factors such as the size and prior attainment profile of a centre's cohort (Rhead, He, et al*.,* 2018; Newton, 2020). Performance differences may be inauthentic, however, if assessments differed in demand or administration context (for example, if performances were produced without the usual time restrictions), or if approaches to aggregation differ with respect to the balance of assessment objectives used to determine a grade.

Teachers were expected to make consistent and reliable judgements based on potentially sporadic student performances gathered for a different purpose. Making reliable summative assessment decisions is a challenging task for teachers (Harlen, 2005), and Wales' Independent Review (Casella et al., 2021) found this to be the case in 2020. They noted that the majority of teachers found that there was sufficient guidance to support judgement, but that the school and college leaders they interviewed disagreed.

Argument 1: The standardisation models

In normal years, the warrant for the awarding phase is primarily focused on comparability between series. The standardisation models were an attempt to meet this warrant using different forms of evidence, and thus apply the same balance of fairness definitions. Neither the *Standard Model* nor the *Banked Unit Model* downgraded students directly. Comparability in standards between centres was a concern for stakeholders (Association of School and College Leaders, 2020), regulator (Qualifications Wales, 2020g) and government (Welsh Government, 2020d) alike, each supporting the need for standardisation in principle. As described in Section 1, given regulatory concerns at the likely quality of the CAGs, and therefore in the absence of strong examinee-level evidence to grade, the standardisation models identified a grade distribution for each centre, to replicate statistical aspects of the systemic definition of standards used in normal years at the cohort level.

Arguments 2 and 3: Centre assessment grades

By removing the standardisation process, a different balance of principles emerges. A CAGs-only model was a highly risk-averse approach towards *aspects* of consequential and retrospective senses of fairness – no student should receive a grade that is below their underlying level of attainment as determined by their teacher, even if some of those grades represented optimistic or unjustifiably generous judgements (Casella et al., 2021). The consequences beyond the immediate progression step (e.g., progression to preferred university placement) were, as will be set out in the next section, less of a concern.

The CAGs-only approach overcame the risk created by standardisation models, of a violation to a central tenet of assessment which relates to social justice at the individual level, that any student could overcome their personal circumstances by performing in their assessments (Baird, 2021) – reflecting both the legitimate expectations and the value-neutral retributive sense. Standardisation models risked preventing anyone from benefiting from performance in their course of study, by not using CAGs (at least in the *Standard Model;* the *Banked Unit Model* did involve the use of student marks from prior series in the qualification being awarded).

No weighting was given in a CAGs-only approach to outcomes at the cohort (examination system) or systemic level. Used alone, CAGs represented an examinee-level definition of standards, in which teachers and senior school leaders were conferred power to contextualise performance and infer to a grade, with the basis for the backing being that they knew the students and their capabilities best (NAHT Cymru, 2020). This position was more commonly held by parents, carers, and learners than teachers and school leaders (Casella et al., 2021). This is a different backing to that for the standardisation models; the question arises if the warrant is the same.

Two separate arguments were provided with respect to the quality and interpretation of grades, representing two different validity arguments for CAGS. The first posited that grades were generous

because of optimism from teachers as to the level of attainment students might have achieved (Holmes et al*.,* 2021), making the overall distribution of CAGs unrealistic (Qualifications Wales, 2020g). This argument for awarding CAGs prioritised optimistic examinee-level judgements based on teacher observations of attainment over causal or cohort comparability. The second stated that genuine attainment is not reflected in examination series; just as teachers could not account for the fact that some students might perform unexpectedly poorly on the day of an exam, exams are less valid than holistic teacher judgement, which can account for all attainment demonstrated over a course of study (Casella et al., 2021). Although in normal years 'special considerations' adjustments are made to marks to account for some events that disrupt performance (JCQ, 2020c), for this argument, the awarding warrant – that the attainment standard is maintained – is qualitatively different because it is argued to rely on a more valid measure of attainment than exams could ever provide. The exam system was said to systematically downgrade students, and thus the higher outcomes seen in 2020 actually more accurately represent student attainment (Harris, 2021).

If one concludes that the CAG grading judgements were sufficiently reliable – that another teacher would have likely awarded the same grade using the same evidence, having accounted appropriately for authenticity concerns relating to the assessment conditions and tasks – then a CAGs-only position can maintain aspects of comparability between students, centres, cohorts, and nations. The associated relational and retributive fairness concerns are also resolved. If reliability is not sufficient, or if there is a more general systematic bias in the inference from performance to attainment, comparability must be downgraded in importance for the CAG-only approach to be considered sufficiently valid. Specifically, inter-centre comparability, the legitimate expectation of previous students that their grades would continue to represent an equivalent level of attainment to the grades awarded in future, and the relational sense of fairness must also be considered less important.

Some advocates for accepting CAGs without standardisation did propose external moderation as an alternative to prevent grades unsupported by performance evidence from being awarded, thus aiding inter-centre comparability (Casella et al., 2021; Holland, in Children, Young People and Education Committee, 2020a). Putting aside the practicalities of undertaking such an exercise when schools were closed, however, moderation would have required a consideration of assessment conditions to ensure comparability in attainment standards. Wales' Independent Review did not conclude that external moderation would have been feasible in practice, suggesting an alternative approach involving professional conversations with Heads of Centres that had approved a profile of CAGs that were far higher than the grades achieved by their centre in previous years (Casella et al., 2021).

Consequences

Reflecting Messick's (1995) concept of consequential validity, the final phase relates to the impact of the consequences of decisions made with the awarded grades. Newton and Shaw's (2014) framework distinguishes between the primary decision-making objective and secondary impacts, as well as distinguishing between technical and social evaluation. The most-discussed *secondary objective* of general qualifications in a normal year is the use of grades in school performance measures that underpin governmental accountability regimes, but these were suspended for 2020 in Wales (Welsh Government, 2020e), and for 2020 in England (Department for Education, 2020d). For this element of this study, therefore, only the primary objective is considered.

The *primary (decision-making) objective* is learner progression. The warrant for the inference is that the information conveyed by the grades is sufficiently useful to appropriately facilitate progress to the next stage of education and/or employment, as is the stated purpose of general qualifications. Two other considerations are vital to consequential validity. Firstly, the awarding process should not lead to too many or too few learners being eligible to progress to scarce opportunities. For those uses, the relative position of learners is as important as the absolute level of attainment reached.

Relatedly, it is important that assessment means that the most appropriate learners are selected for courses that they are able to complete, or undertake jobs they are capable of – this is predictive validity.

For Newton and Shaw, the social evaluation of the primary objective must cover four concerns, which overlap. Credibility is drawn from stakeholder views of the assessment process and its outcomes. Legality relates to the extent to which the process is compliant with the law. An evaluation of utility requires an economic cost-benefit analysis of the process. This has not been undertaken rigorously with respect to UK general qualifications, although some voices have in recent times attempt to estimate the cost of the process as a means of questioning its value (Richmond, 2021).

Finally, fairness is considered a question for social evaluation. This includes considerations more clearly in line with the key principles set out above, especially comparability and fairness. People with grades from different examination series (or subjects) will want reassurance that their grades have the same meaning in the world as those achieved by other candidates, within the relational sense of fairness. Candidates in different groups will expect to be treated the same through the assessment process, in the formal sense of fairness. Most obviously, the consequential sense of fairness arises when decisions are made using candidates' grades.

Newton and Shaw's framework also covers the *broader impacts* of assessment policy. Newton and Shaw illustrate this with reference to learners who do not achieve the benchmark grade (4 or higher) needed at GCSE to progress to the next level of education. Grades within a single scale can have differing purposes (Newton, 2017), some being associated with 'gate-keeping decisions' and others being progression markers that represent a level of attainment without have much purpose beyond that. This is therefore where the technical and social aspects overlap: if some grades matter more than others because of the decisions that are made using them, it may be necessary to focus on the reliability (and by extension the classification accuracy) around grading for those key grades.

Crooks, Kane and Cohen (1996) also suggest that creating unmanageable stress or damaging the relationships between candidates also reduces validity.

Consequences in 2020

The broader impact of the standardised grades was illustrated by the volume of press coverage and disquiet that standardisation had caused; the higher level of public confidence in CAGs, over standardised grades, became clear shortly after AS and A-level results day in Wales. The planned use of an appeals mechanism to reduce the impact of the standardisation model on outlying cases was also swiftly undermined by the limited grounds for appeal (Plaid Cymru, 2020). Conversely, Wales' Independent Review and Ofqual's research (Casella et al*.,* 2021; Holmes et al., 2021), teachers expressed concerns that some students' grades were in some cases inflated and might not have been suitable for progression, although this appears to have been a minority view.

The standardisation models used in Wales were ultimately overturned to ensure relational fairness by protecting inter-jurisdictional comparability (BBC, 2020b). For A-levels, progression to university was disrupted for many learners once standardisation was rejected, requiring flexibility from universities and the need to overcome resource constraints (Blaker, in Children, Young People and Education Committee, 2020b). Views on this disruption differ; although there have been consequences for courses such as medicine in the UK (Medical Schools Council, 2021), the Independent Review noted that disruption was limited to universities not having advance notice of the policy change, noting that there was too much emphasis placed on managing the scarcity of places on the most competitive courses and not enough on lower-attaining students.

This argument rests of the balance of importance between grades recognising a given level of attainment – an *absolute* judgement for each student – and the need to ensure that qualifications meet their purpose in informing selection – a *relative* judgement which requires differentiation between students. This distinction was recognised in normal years via rank orders based on

common assessments and mark scales, and the standardisation process via CAGs and rank orders; CAGs alone represent only an absolute judgement.

Balancing key principles in a validity argument

The standardisation models, CAGs and a normal examination series have different validity arguments for, notionally, the same qualifications. The standardisation models emphasised that meeting the purposes of general qualifications required the same dual focus on the examinee and cohort levels when maintaining attainment standards, even in the absence of common assessment processes. Standardisation was used to overcome any issues with the reliability and comparability of awarding inferences, which could be caused by unreliable teacher judgement, differences in assessment conditions and tasks.

The validity argument for CAGs challenges the idea that you can (or should) maintain a similar balance of key principles of assessment to a normal year. In the absence of controls that enable an exam board to identify what is fair for every single examinee, validity for grading judgements ought to be delegated to teachers, even if this requires a trade-off involving other principles, in particular the relegation of comparability between cohorts and centres, and relational fairness. It is better to give everyone the grade that teachers (in whom power is conferred) judge they 'deserve' (the evaluative form of retributive fairness) than risk damaging that by accounting for concerns around reliability of grading judgements, comparability and relational fairness.

While the profile of CAGs was higher overall than the grade profile seen in the same qualifications in 2019, the question remains as to how this should be interpreted. Wales' Independent Review claimed that no assessment of the extent of any 'grade inflation' could be made, as no review of centre submissions deemed atypical compared with historical data trends was made (Casella et al. 2021). Given that the validity argument of using CAGs as grades in 2020 rests in part on the

likelihood of them representing the levels of attainment of the students who were ultimately awarded them, however, the next section of this study focuses on this aspect.

Section 4: Assessing the quality of centre assessment grades: the case of Wales A-levels

Introduction and research questions

The question that teachers were asked to answer when submitting their CAGS in 2020 appears a straightforward judgement at first glance: what grade would the student most likely have achieved had exams taken place? (Qualifications Wales, 2020e). Even the question, however, shows that this judgement relates to an outcome that would have been uncertain even if normal assessment procedures had been completed. Although this element was not designed to form the sole basis for awarding grades, there are two reasons to frame this question in terms of uncertainty. Firstly, no assessment is perfectly reliable. Secondly, answering the question requires two cognitive steps – a judgement of each student's level of attainment prior to the pandemic, and an extrapolation from that judgement as to how much more knowledge, skills and understanding the student would have attained had teaching and learning have continued as normal until examinations had taken place.

Other research has attempted to set out how patterns of attainment in 2020 compare with previous years for groups with different personal characteristics, as well as for centre types (Qualifications Wales 2020h). The purpose of the analysis presented here is different – to establish the plausibility of the CAGs provided by centres, assessed on the terms upon which they were collected: the likelihood of students achieving those grades, based on their prior attainment on each qualification.

Whereas general qualifications in England have a linear structure, with all assessments being completed at the end of the course of study, in Wales A-levels are unitised. Units can be entered each summer; AS units are usually completed at the end of the first year of study, and students usually use these to 'cash-in' an AS qualification grade. A2 units – usually of higher demand – are completed at the end of the second year, with outcomes from AS units (worth 40%) and A2 units (worth 60%) combined to form the basis for the final A-level qualification grade. Students may also

resit AS units to improve their A-level grades. Alternatively, they may choose not to progress to A-level.

Given the high-stakes nature of the AS assessment and their nesting within the A-level assessment structure, we might expect that AS grades and marks would, in a normal year, represent a strong measure of attainment halfway through each student's course, and be a good predictor of A-level grades. If this relationship holds, historical data on Wales AS and A-level outcomes can establish the probability of standardised grades, CAGs and final grades awarded to A-level students in 2020, and the extent to which higher outcomes overall were due to either the award of implausible CAGs given students' prior attainment, and/or an issue of scale – that is, if plausible CAGs at the examinee level scaled to a very unlikely outcome at the cohort level.

At the end of the section, the findings are discussed with reference to the key principles of assessment set out in Section 3, as well as the critiques of each approach to grading.

Method

Data set

In each examination series, students' raw marks are converted to uniform mark scale (UMS) values to ensure that the scores achieved by each student are comparable from year to year. For all subjects, the maximum possible total UMS is 200 for AS units, 300 for A2 units, and 500 for A-level overall. The exceptions are Mathematics and Further Mathematics (240 for AS units, 360 for A2 units, so 600 for A-level). To prevent this inconsistency from affecting the analysis, each student's total AS UMS value for each subject was converted to a percentage so that all values were on the same scale.

The data set used was limited to students who a) either achieved an AS grade in year one of their courses or completed all relevant units, and b) either achieved a valid A-level grade in 2019 or

received a valid CAG in 2020. To improve the comparability of the data between the two years under investigation, each data set was limited to students registered to centres in Wales aged 18 on 31st August in the academic year in which the A-level was achieved. In addition, private candidates, who may not have a teaching relationship with their exam centre, were removed from the analysis.

The data file for analysis was compiled by WJEC staff to the specification set out in *Table 4* and was based on data extracted from WJEC's data store on 27th April 2021. Students were not identifiable in this data set. As shown in *Table 5,* the 2019 and 2020 cohorts were comparable in terms of size and profile of banked AS results (graded/ungraded, intention to re-sit units and mean grade performance).

Table 4: Variables in final data set

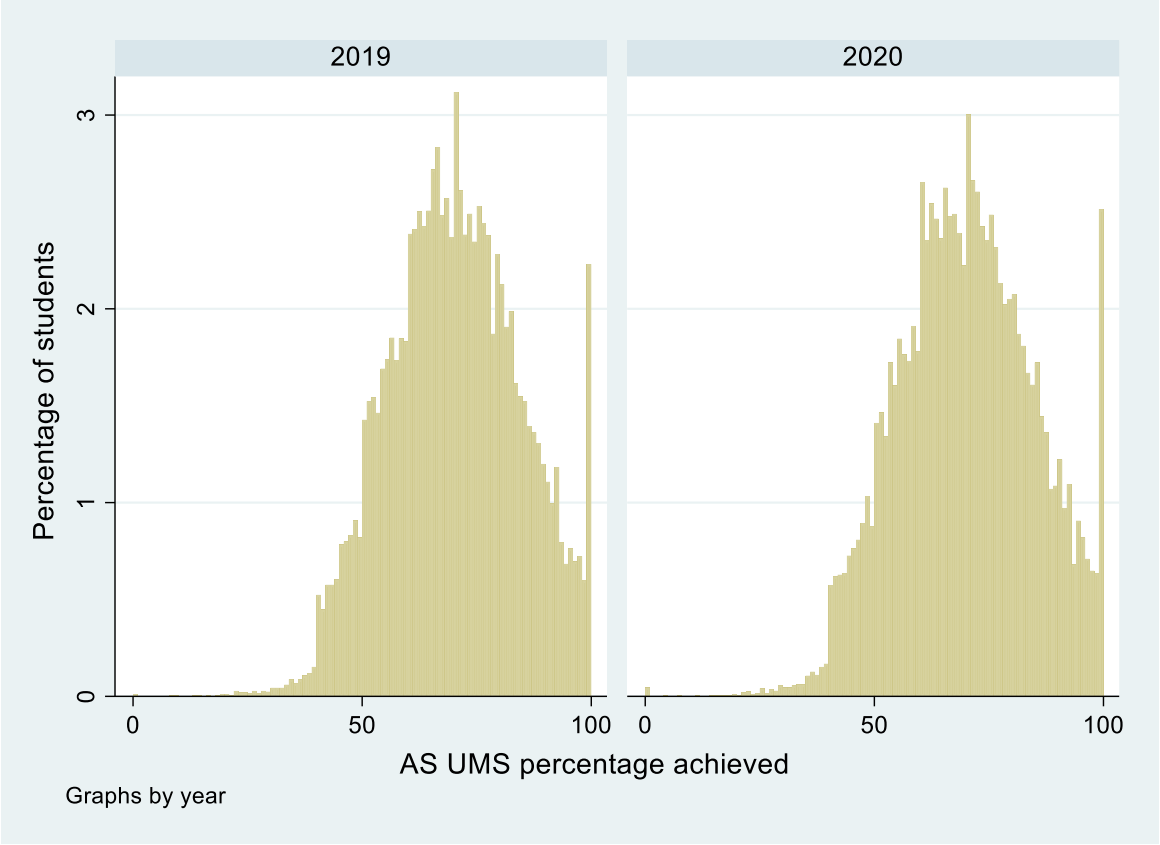
Variable name	Description	Type (minimum, maximum)	Number of observations
ASgrade_num	AS grade	Ordinal (U=0,A=5)	43,388
ASgrade_num_inf	AS grade including inferred grades	Ordinal (U=0,A=5)	49,392
ASperc	Total AS UMS percentage score	Scale (minimum=0, maximum=100)	49,392
Alvlgrade_num	Final A-level grade	Ordinal (U=0,A*=6)	49,392
StandardisedGrade_num	Standardised grade (2020 only)	Ordinal (U=0,A*=6)	24,379
CAG_num	CAG (2020 only)	Ordinal (U=0,A*=6)	24,379
Alvlperc	Total A-level UMS percentage score	Scale (minimum=0, maximum=100)	25,013
Resit	Number of AS units entered as a resit in the year that the A-level was cashed-in	Scale (minimum=0, maximum=3)	49,392
Year	Year in which A-level grade was achieved: 2019 or 2020	Ordinal	49,392 (2019: 25,013; 2020: 24,379)

Table 5: Description of data set

	Year in which A-level grade was achieved	
	2019 (n)	**2020 (n)**
Total number of A-level grades awarded to Wales 18-year-olds (excluding private candidates)	25,429	24,825
Total number of A-level grades with AS performance available for comparison in the analysis *(percentage of all grades awarded)*	25,013 *(98.4%)*	24,379 *(98.2%)*
- Number of A-level grades with an AS grade associated with it from the previous year	21,899	21,489
- Number of A-level grades with no AS grade associated with it from the previous year, but with all AS units completed	3,114	2,890
Total number of grades for which AS units were to be re-sat in the second year of study *(percentage of all grades awarded)*	9,718	9,696
Mean AS grade (standard deviation) in previous year U=0, E=1, D=2, C=3, B=4, A=5	3.45 (1.26)	3.41 (1.29)
Mean A-level grade (standard deviation) U=0, E=1, D=2, C=3, B=4, A=5, A*=6	3.55 (1.42)	3.61 (1.40) *(standardised grades)* 4.09 (1.24) *(CAGs)* 4.20 (1.19) *(final grade)*

The distribution of AS UMS percentages by year is shown in *Figure 4*. In both years, scores largely follow a normal distribution, but with very few scores below 40% (the cut-off for a grade E at AS-level) and a significant number achieving 100%. The latter represents a ceiling effect and occurs because of the way UMS is calculated for each unit in each examination series. The proportion achieving the maximum is relatively low at under 3% of the cohort in each year and therefore no adjustment was made to account for this in the analysis.

Figure 4: Distribution of AS UMS percentages by year of A-level result



Year	Mean	Standard deviation
2019 A-level cohort	69.94	14.45
2020 A-level cohort	69.73	14.84

Table 6 sets out the cumulative grade distribution for the 2019 and 2020 A-level cohorts. This includes inferred grades for those with banked AS UMS but who did not cash-in that qualification in their first year of study. Including inferred grades does not have a large effect on overall AS outcomes; no difference exceeded one percentage point. 2020 standardised A-level cohort outcomes were higher than in 2019; 2020 CAGs and the final grades awarded following the post-results change in policy were higher still.

Table 6: Cumulative grade outcomes (Wales 18-year-olds)

	A*	A	B	C	D	E	U	*n*
2018 AS	n/a	26.1%	50.7%	76.1%	92.6%	99.3%	100.0%	100.0%
2018 AS including inferred grades	n/a	25.8%	50.2%	75.4%	92.1%	98.9%	100.0%	100.0%
2019 A-level	8.8%	26.5%	52.4%	76.9%	92.0%	98.3%	100.0%	100.0%
2019 AS	n/a	25.6%	50.0%	74.9%	91.5%	98.9%	100.0%	100.0%
2019 AS including inferred grades	n/a	25.9%	50.0%	74.6%	91.1%	98.7%	100.0%	100.0%
2020 A-level – *standardised*	9.4%	27.2%	54.1%	78.5%	93.0%	98.6%	100.0%	100.0%
2020 A-level – *CAGs*	14.1%	39.0%	67.7%	90.1%	98.1%	99.9%	100.0%	100.0%
2020 A-level – *final grade*	15.4%	42.0%	71.2%	92.5%	99.0%	100.0%	100.0%	100.0%

Model selection and analytical approach

Ordinal logistic regression was used to create a proportionate-odds model of the relationship between the total UMS achieved at AS in 2018 (independent variable) and A-level grade in 2019 (dependent variable), using the *ologit* command in the statistical analysis software package Stata 17 (StataCorp, 2021). This form of regression was chosen because grade data is both categorical and strictly ordered: a grade A covers a range of UMS but is always ranked higher than a grade B. This approach is used to analysis grade outcomes (*cf*. Yen & Liu, 2009) as it requires no assumptions to be made about the distance between grades being equal – that is, the grade scale is not quantitative – but nevertheless optimises analysis focused on distinctions between ordinal categories (Harrell, Jr., 2015).

An ordinal logistic regression (OLR) analysis was conducted to confirm that a student's A-level grade could be predicted from the UMS percentage they achieved at AS. Only one predictor variable was used, so there could not be a violation of the assumption of multicollinearity. The predictor variable in the OLR analysis, AS UMS percentage, was found to contribute to the model. The log-likelihood ratio Chi-Square test with 1 degree of freedom, LR $\chi^2 = 26792.11$, $p<.001$ indicated that the model provided a better fit than the null model with no independent variables.

The estimated logit regression coefficient, $\beta = 0.184$ [95% CI (0.181, 0.187)], SE=0.001, $z = 132.41$, $p < 0.001$, indicated that the AS UMS percentage had a significant effect on students' grades. McFadden's pseudo-$R^2 = 0.3070$, suggested that the model was a good fit (McFadden, 1979). Expressed as an odds ratio, 1.202, this indicated that odds of being at or below a particular grade relative to beyond that level decreased by a factor of 1.202 with a one-unit increase AS UMS percentage.

The Brant test of the proportional-odds assumption was completed using Stats *brant* command (Long & Freese, 2014). This indicated that the effect of increasing AS UMS percentage was not similar across the grades (χ^2=142.95, p<.001); thus, the assumption was violated. A generalised OLR approach, using the *gologit2* command in Stata to relax the assumption, was also calculated. In this model, the log-likelihood ratio Chi-Square test with 1 degree of freedom, LR $\chi^2(6) = 26911.06$, p<.001 indicated that the model provided a better fit than the null model with no independent variables. McFadden's pseudo-$R^2 = 0.3084$, suggested that the generalised OLR model was a similar, but slightly higher, level of fit than the OLR model. As shown in *Table 7*, the estimated logit regression coefficients were significant for all grades (p<.001), again indicating that the AS UMS percentage had a significant effect on students' grades. *Table 8* shows that both model selection statistics – the Akaike information criterion (AIC) and Bayesian information criterion (BIC) – were lower for the generalised OLR model. The generalised OLR model was therefore used for the analysis presented in the Results section.

Table 7: Parameter estimates for generalised ordinal logistic regression model

A-level grade	Variable	Coefficient β	Odds ratio	Std.error	z	P>[z]	95% confidence interval	
U	AS UMS %	0.192	1.211	0.008	29.140	0.000	1.196	1.227
	constant	-6.650	0.001	0.000	-21.110	0.000	0.001	0.002
E	AS UMS %	0.174	1.190	0.004	51.780	0.000	1.182	1.198
	constant	-7.890	0.000	0.000	-43.520	0.000	0.000	0.001
D	AS UMS %	0.176	1.192	0.003	72.750	0.000	1.187	1.198
	constant	-10.024	0.000	0.000	-67.690	0.000	0.000	0.000
C	AS UMS %	0.181	1.198	0.003	81.800	0.000	1.193	1.203
	constant	-12.467	0.000	0.000	-81.030	0.000	0.000	0.000
B	AS UMS %	0.204	1.227	0.003	79.510	0.000	1.221	1.233
	constant	-16.591	0.000	0.000	-81.350	0.000	0.000	0.000
A	AS UMS %	0.177	1.193	0.004	58.950	0.000	1.186	1.200
	constant	-16.680	0.000	0.000	-63.410	0.000	0.000	0.000

Table 8: Model selection statistics

Model	N	ll(null)	ll(model)	df	AIC	BIC
OLR	25,013	-43628.6	-30232.6	7	60479.12	60536.01
Generalised OLR	25,013	-43628.6	-30173.1	12	60370.17	60467.69

The equations from the regression analysis were then applied to the total AS UMS achieved in 2019 towards each 2020 A-level grade, thus obtaining an estimated probability of each student achieving each of the seven outcomes (in increasing order of attainment – U, E, D, C, B, A, A*) in 2020. These could then be compared with the CAGs submitted by centres and the final grades awarded to students, to answer the research questions.

Crosstabulations of various statuses of achieved AS and A-level grades were produced to identify changes in grades through the assessment and grading processes. Descriptive statistics, boxplots and histograms were used to visualise differences in probability distributions for the various grade types (standardised grades, CAGs, and final grades).

Finally, grade probabilities were used to simulate cohort-level grade distributions. 10000 simulations were run. For each simulation, every grade was assigned a randomly-generated value between 0 and 1, based on a uniform distribution. The cumulative probability for each grade was compared with the random value; the highest grade with a probability equal to or higher than the random value was awarded.

Results

Figure 5 illustrates the relationship between the AS UMS percentage achieved in 2018 and the final A-level UMS percentage in 2019. Logistic regression curves are also shown. Even the maximum AS UMS percentage does not guarantee an A* at A-level, reflecting the fact that AS is only worth 40% of the total qualification. At the opposite end of the grade scale, relatively few students with an AS UMS percentage below 40% in 2018 continued to A-level in 2019. As a result, as can be seen in the graph, a wide range of AS UMS percentages are associated with each A-level grade.

Figure 5: Relationship between AS UMS percentage and A-level UMS percentage in 2019, by A-level grade achieved, with generalised ordinal logistic regression curves by A-level grade

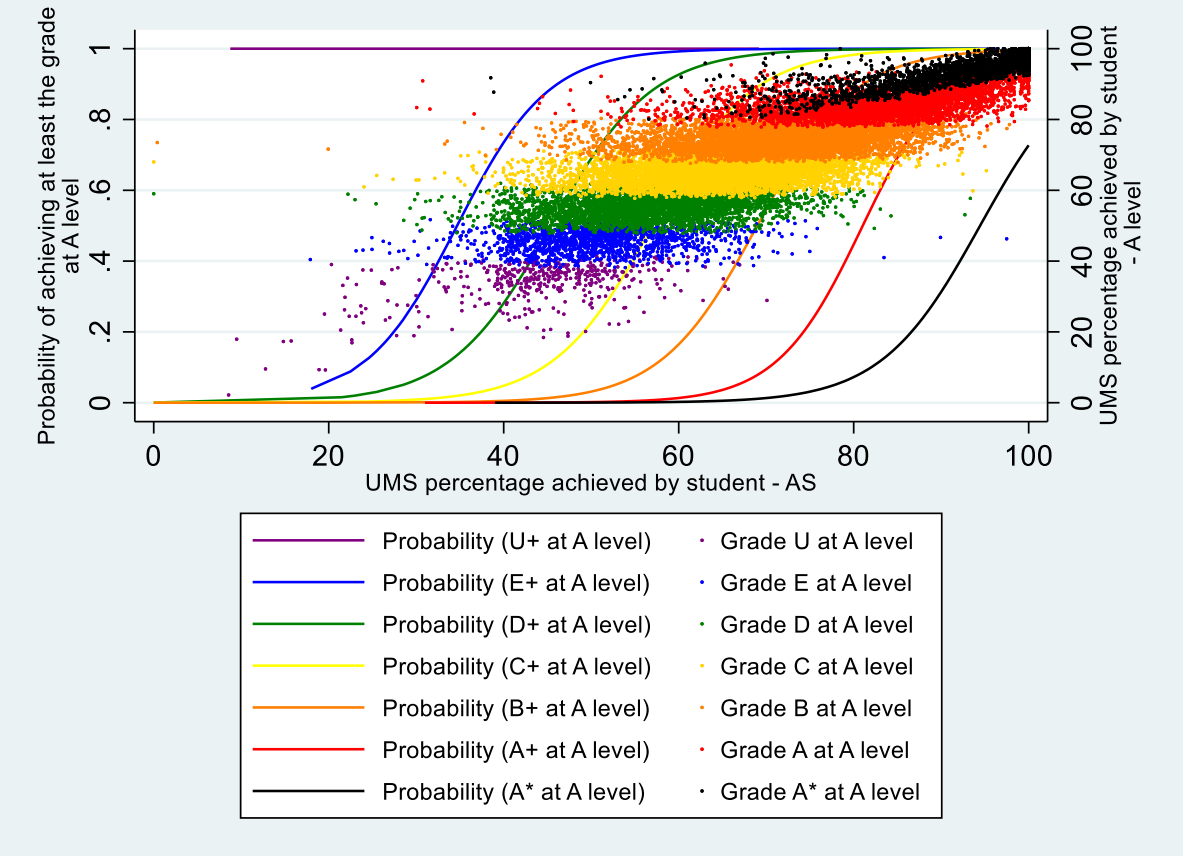


Figure 6 shows the boxplots and descriptive statistics for the distribution of AS UMS percentage scores by A-level grade, for 2019 and for each of the three grade types awarded in 2020. For all four grade sets, the mean AS percentage increased with improving A-level grades. For the standardised grades, the means at grades A, B and C were similar to 2019, although the mean was higher at A*

and somewhat lower at grades D and E. The mean scores for the CAGs and the final grades were, with the exception of U, all below those for 2019 and the 2020 standardised grades. The very low number of U CAGs and final grades awarded in 2020 make comparisons less meaningful. At A*, the mean AS UMS percentage is very similar for CAGs and the 2020 final grade to that found in 2019.

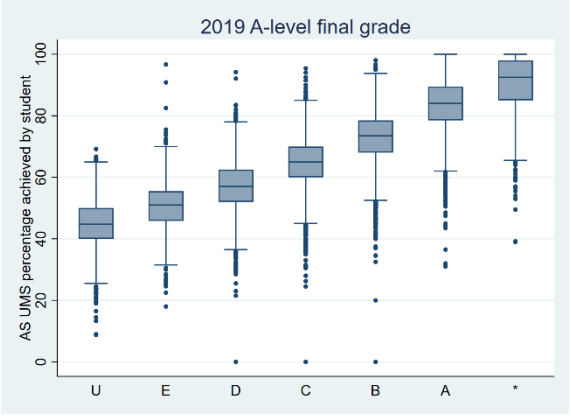
Table 9 shows how the relationship between AS and A-level grades differed depending on the A-level grade status. Even for the 2020 standardised grades, students were less likely to achieve a lower grade in their A-level than their AS compared with 2019 (22.4% in 2019; 11.0% in 2020 for standardised grades); the proportion achieving a higher grade was similar to the previous year (31.2% and 30.1%, respectively). Far fewer achieved a lower A-level CAG than AS grade (4.4%), while – due to the grading rule introduced post-results – no student achieved a lower final A-level grade in 2020 than their AS grade. More than half of CAGs (59.2%) and final grades in 2020 (62.9%) were higher than the preceding AS grade.

Table 9: Difference between A-level grade and AS grade in the previous year*

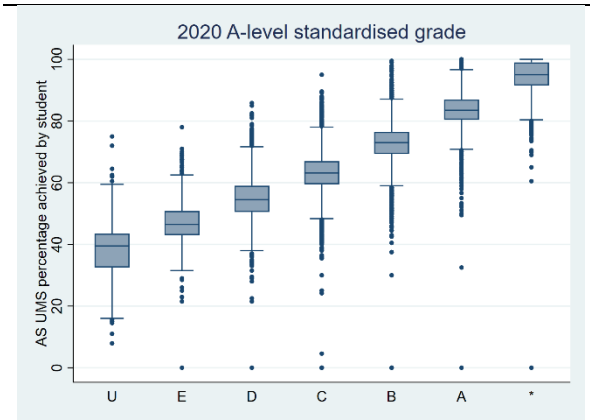
	Final A-level grade (2019)	Standardised grade (2020)	CAGs (2020)	Final A-level grade (2020)
AS grade > A-level grade by two or more	2.7%	1.0%	0.5%	0.0%
AS grade > A-level grade by one	19.7%	10.1%	4.0%	0.0%
AS grade and A-level grade are equal	46.4%	58.9%	36.4%	37.1%
A-level > AS grade by one	25.7%	27.8%	46.8%	48.8%
A-level grade > AS grade by two or more	5.5%	2.3%	12.4%	14.1%

* Includes grades inferred from AS UMS awarded.
n=25,013 (2019), n=24,379 (2020)

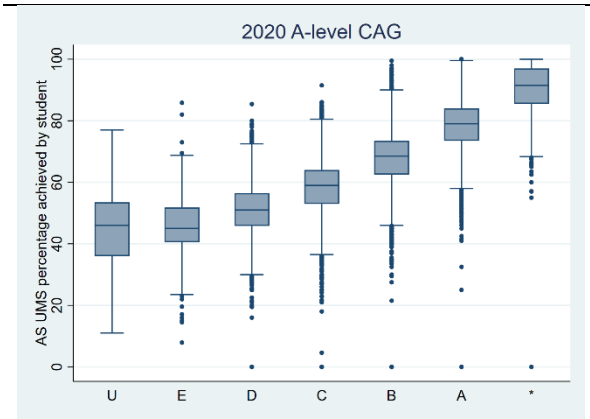
Figure 6: Distribution of AS UMS percentages, by A-level grade in the following year



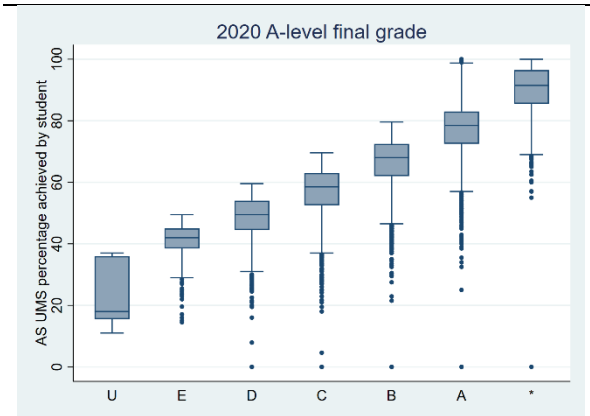
Grade	n	Mean	Median	St.Dev
U	436	44.0	44.8	9.5
E	1564	51.0	51.0	7.9
D	3775	57.0	57.0	8.2
C	6137	64.6	65.0	8.3
B	6479	72.7	73.5	8.4
A	4417	83.3	84.0	8.8
A*	2205	90.5	92.5	9.0
Total	25013	69.9	70.0	14.5



Grade	n	Mean	Median	St.Dev
U	353	38.5	39.5	9.7
E	1343	47.4	46.5	6.7
D	3535	54.9	54.5	7.0
C	5960	63.0	63.2	7.0
B	6548	72.5	73.0	6.5
A	4349	83.4	83.5	6.2
A*	2291	94.2	95.0	5.7
Total	24379	69.7	69.6	14.8



Grade	n	Mean	Median	St.Dev
U	19	45.2	46.0	18.7
E	440	45.5	45.0	9.9
D	1955	51.4	51.0	8.8
C	5469	58.4	59.0	8.5
B	6990	67.5	68.5	8.6
A	6060	78.4	79.0	8.5
A*	3446	90.6	91.5	7.5
Total	24379	69.7	69.6	14.8



Grade	n	Mean	Median	St.Dev
U	5	23.5	18.0	12.1
E	241	40.6	42.0	6.8
D	1586	48.7	49.5	7.1
C	5180	57.2	58.5	7.6
B	7132	66.6	68.0	8.2
A	6492	77.4	78.5	8.7
*	3743	90.6	91.5	7.4
Total	24379	69.7	69.6	14.8

Figure 7a shows the distribution of the probability of the *most likely* grade for each A-level awarded in 2019 and 2020, based on the model and student performance at AS; the distributions are very similar. In 2020, for 70.6% of grades, the most likely A-level grade had under a 50% chance of being achieved, based on the relationship between AS performance and A-level grades in 2019. *Figure 7b* shows the relationship between the probability of the most-likely and second most-likely A-level grade, for each 2020 A-level grade awarded; the second most-likely grade usually has a probability of between 0.2 and 0.5.

Figure 8 shows the modelled probability and cumulative probability of each A-level grade awarded: 2019 final grades, 2020 standardised grades, 2020 CAGs and 2020 final grades. In 2019, the most common probability was the one approaching certainty, in part reflecting the higher number of U grades awarded in that year, compared with 2020. The distribution of cumulative probabilities is skewed to the left. The standardised grades distribution is somewhat similar, but with a higher proportion of students achieving the 'most likely grade' and fewer grades at either end of the grade distribution. The distribution of probabilities for CAGs is somewhat different, with far more students achieving grades with a relatively low probability; this means the cumulative probability distribution is much flatter than in 2019 and the 2020 standardised grades.

Figure 7a: Highest modelled grade probability for A-level grades, 2019 and 2020

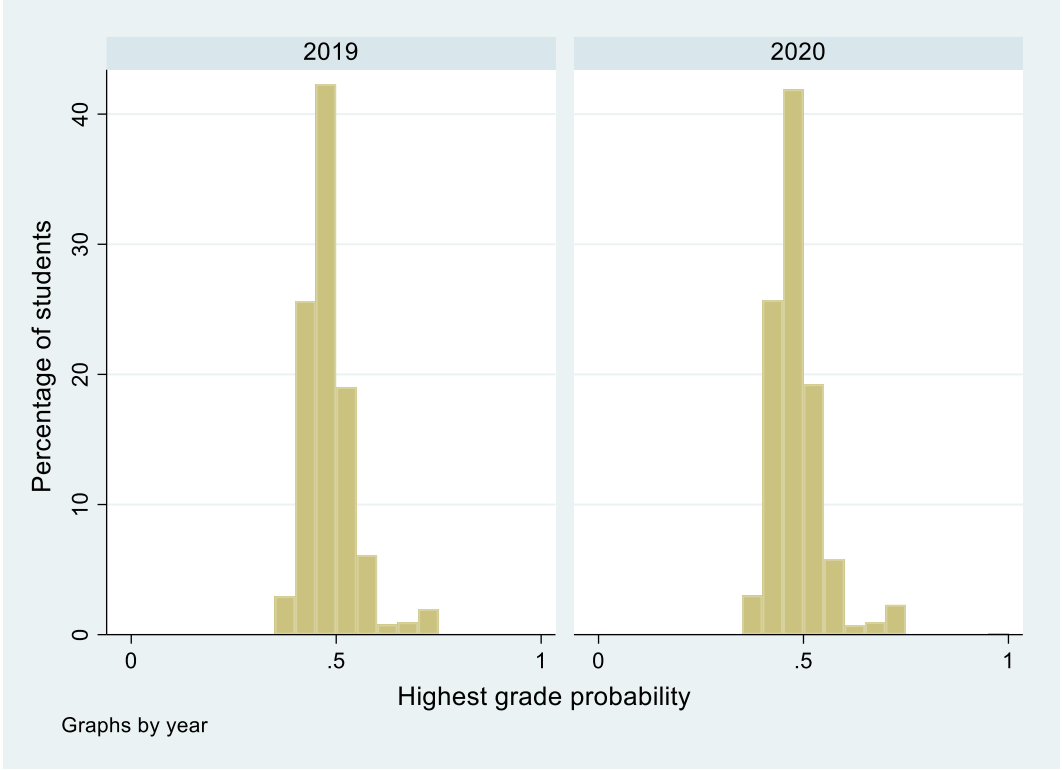


Figure 7b: Highest and second-highest modelled grade probability for A-level grades, 2020

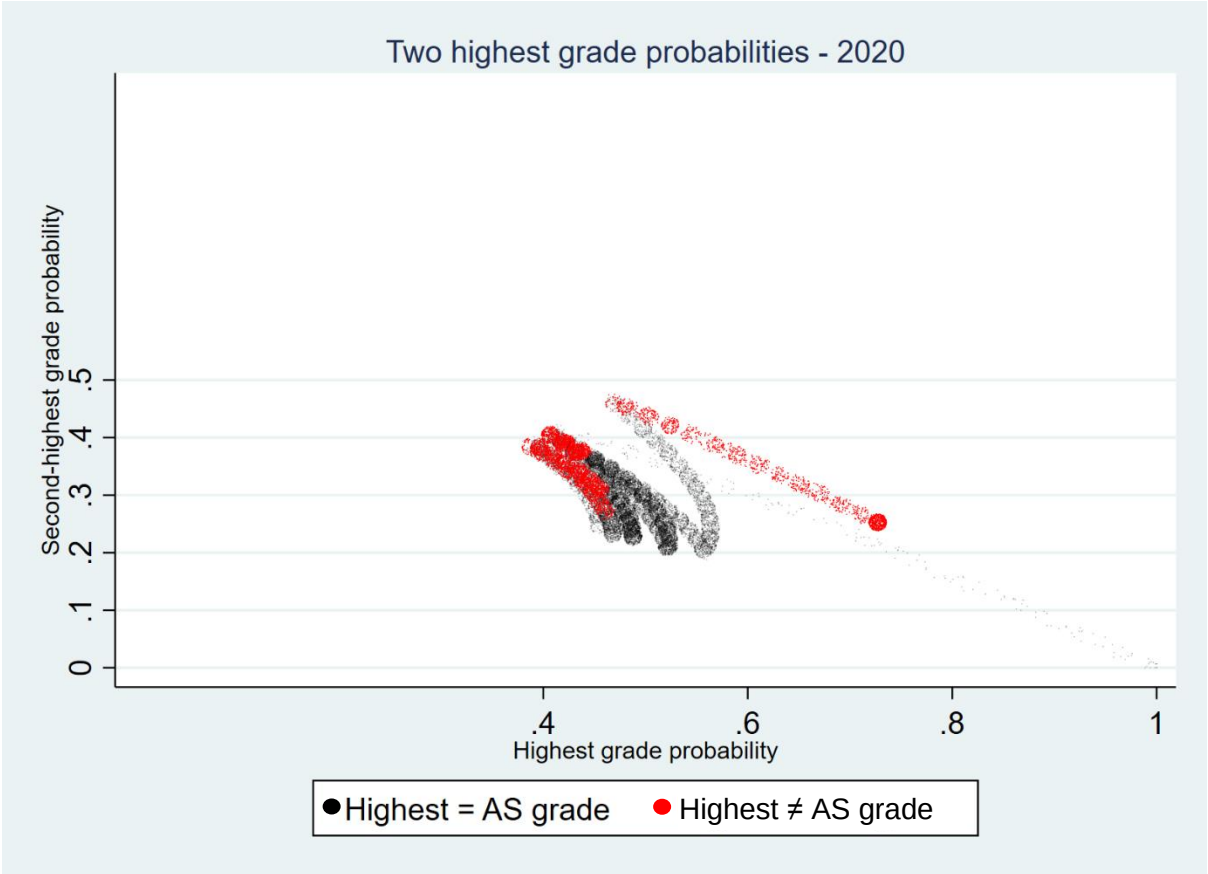


Figure 8: Modelled probability and cumulative probability of each A-level grade awarded: 2019 final grades, 2020 standardised grades, 2020 CAGs and 2020 final grades

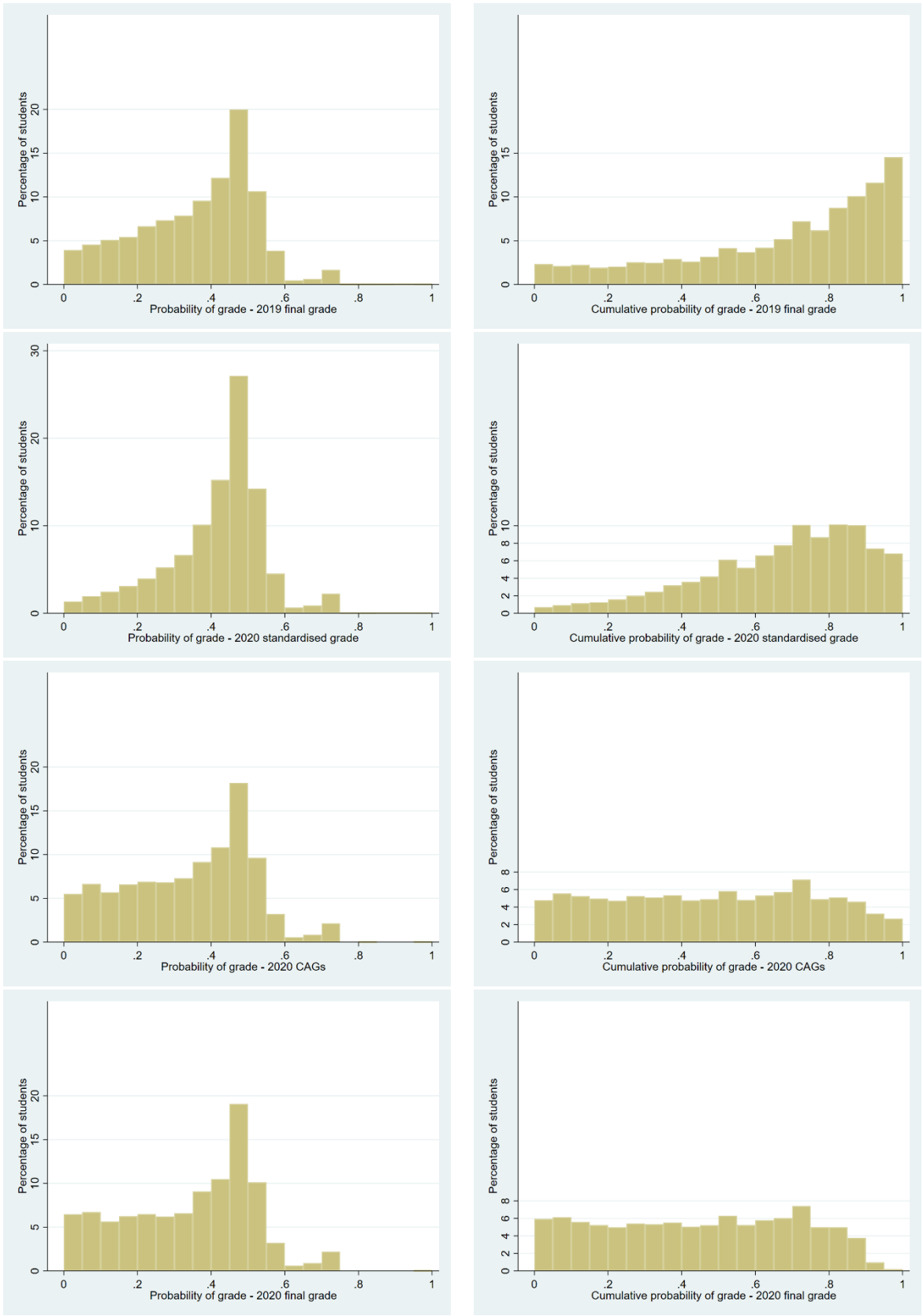


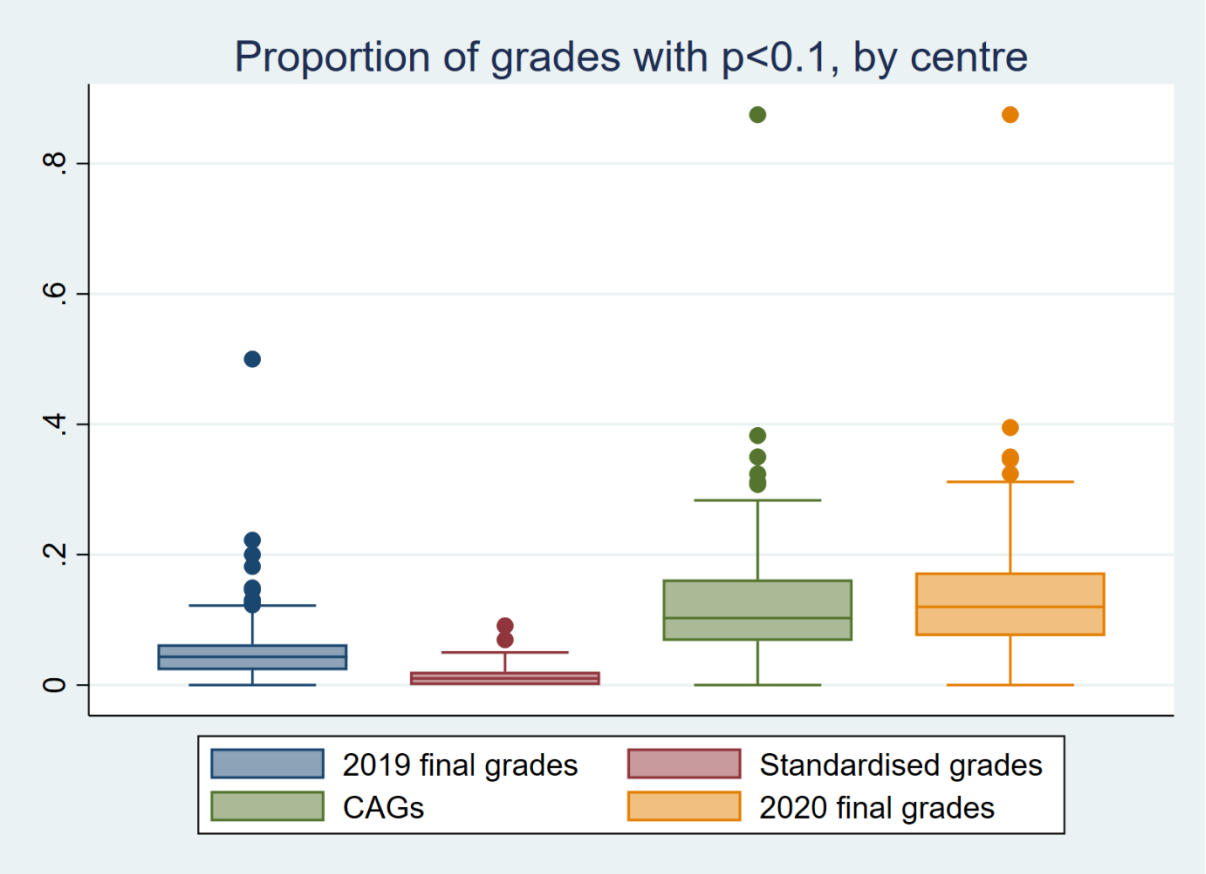
Table 10 shows the difference between the grade awarded (by grade status) and the most likely grade. The standardisation model led to a higher proportion of the 'most likely' grade (according to the generalised OLR model) being awarded than in 2019, with very low occurrences of large divergence. Almost as many CAGs were awarded one grade above the most likely, however, with under 6% awarded below the most likely grade.

Table 10: Difference between 'most likely' A-level grade and awarded A-level grade, by grade status

	2019 final grades	2020 standardised grades	2020 CAGs	2020 final grades	Model effect
Grade awarded was more than one lower than most likely	3.7%	1.4%	0.7%	0.0%	**Overestimate**
Grade awarded was one lower than most likely	24.2%	14.2%	5.8%	1.8%	
Most likely grade was awarded	48.9%	63.2%	43.7%	44.8%	**Accurate**
Grade awarded was one higher than most likely	19.3%	17.2%	38.7%	40.7%	
Grade awarded was more than one higher than most likely	3.9%	1.4%	8.6%	10.1%	**Underestimate**

Figure 9 illustrates proportion of grades awarded with a modelled probability of below 0.1, by centre. In most cases, centres achieved a low proportion of grades with a low likelihood; this was much lower for the standardised grades, but much higher for the CAGs.

Figure 9: Proportion of grades achieved with a probability below 0.1: 2019 final grades, 2020 standardised grades, 2020 CAGs and 2020 final grades



The cohort simulations based on the generalised OLR model are summarised in *Table 11*. The mean outcomes are slightly lower than the cumulative grade distribution in 2019, reflecting the slightly lower mean AS UMS percentage in 2020. Even the highest simulated outcomes at each grade fall close to the 2019 distribution.

Table 11: Summary of simulated outcomes for 2020

Grade	Mean cumulative grade outcome	Standard deviation	Minimum	Maximum
A*	8.76	0.15	8.22	9.39
A	26.11	0.19	25.30	26.80
B	51.49	0.22	50.71	52.30
C	75.75	0.20	75.00	76.55
D	91.14	0.15	90.65	91.73
E	97.82	0.08	97.51	98.11
U	100.00	0.00	100.00	100.00

Discussion

It was known prior to the design of the 2020 grading process that teachers were good at ranking students relative to each other (Baird, 1997) but tended to err in favour of students when predicting their grades (Dhillon, 2005), particularly students with lower prior attainment (UCAS, 2016). Predicting A-level grades based on past performance is known to be very difficult, even using advanced statistical techniques (Anders et al., 2020). To evaluate whether the CAGs were the grades that students would most likely have achieved in a normal year, as intended, we must establish if any, or several, of the options available to teachers when setting each grade would have been considered likely. Given the importance of public confidence in grading processes (Ofqual, 2020d; Qualifications Wales, 2020e; Weinberg, 2020), understanding what the public might consider to be a likely grade can help in interpreting the analysis presented in this section.

In July 2020, the polling company YouGov gathered data on how British adults perceive probability more generally. As noted above, for most students, their most likely A-level grade in 2019 and 2020 had a probability of between 40% and 50%. A 40% chance of an occurrence happening was felt to be likely by less than one-third of adults (Smith, 2020). In a normal year, which side of a boundary a student ended up would relate to the reliability of the assessment, and their performance in the assessment. In 2020, with a year with no standardised assessment, in most cases teachers were faced with a choice between two or more possible grades – one being equivalent to the student's AS grade – and none of which would have had a high likelihood of being achieved has exams have taken place, as *Figure 7b* shows. In many cases, the AS grade was not the highest probability grade, putting teachers in a difficult position – to award a CAG below the grade achieved the previous year, or a more likely but lower grade? The lack of a highly likely grade in most instances reflects the fact that, while the relationship between AS and A-level is strong, lots of factors can affect attainment between the two in the year between AS assessments and A2 assessments (and the determination of CAGs in 2020).

While standardisation of A-level grades didn't quite replicate grade patterns at the cohort level from 2019, with outcomes generally higher overall, the profile of grades – and the probability of those grades occurring in a normal year – is more similar than for CAGs. This is not to say that the probability of each CAG is implausible, however; rather that, when faced with a choice between two possible grades, teachers are likely to have selected the higher grade. In the analysis presented above, 38.7% of all CAGs were one grade higher than the most likely grade, and 8.6% of all CAGs were two grades higher than the most likely grade; the equivalent proportions for underestimates were 5.8% and 0.7%. Research evidence from England suggests, when determining CAGs, in the absence of strong evidence for identifying those students, often all within that group were awarded the higher grade (Holmes et al., 2021; Casella et al*.,* 2021). This accounted for some, but not all the differences. The proportion of very unlikely grades (judged by 80% of adults in the YouGov study as corresponding to lower than a 10% probability) was also notably higher for CAGs, and was found to be across most centres, indicating factors beyond a 'benefit of the doubt' selection in many cases.

The results of the cohort-level simulation indicate strongly that neither the profile of CAGs nor the final grade profile were plausible outcomes given students' previous unit UMS scores. The simulations embody the principle that, in a normal year, some students will perform above expectations in an examination series, but other students performing below expectations will tend to balance that out, unless there are systemic factors in play. In a normal year, this reflects the level of assessment reliability. Within the standardisation model, this principle was applied at centre level: while no centre could break outside of the expected outcome for its cohort, a student within the centre could, by changing their position within the centre rank order. There is a distinct contrast between the proportion of low-probability grades achieved in a normal year, compared with the much lower proportion emerging from the standardisation model, and the much higher proportion emerging from the CAGs. Although standardisation largely prevented the large improvements in grade by students between AS and A-level seen occasionally in normal years, they were far more frequently seen as a result of awarding CAGs.

There are some limitations to the presented analysis. Firstly, the results are limited to A-levels in Wales, whereas the focus of earlier sections is broader geographically. Although differences in outcomes between equivalent linear and modular qualifications have not been found to be educationally significant (Baird et al., 2019), the modular structure used in Wales A-levels was previously found to lead to more accurate teacher predictions (Dhillon, 2005). In this respect, teachers in Wales could be said to be at an advantage when predicting the most likely grade, and so the findings may overestimate the predictability of A-level grades in other contexts. Secondly, because the CAGs were not initially collected to be the final grades, there is a possibility that teachers deliberately sought to err on the positive side, knowing that the grades would be standardised. This behaviour was not identified by the Independent Review in Wales but was referred to in Ofqual's research into CAGs (Holmes et al., 2021).

Thirdly, using the relationship between AS and A-level outcomes to assess the different approaches to grading risks replicating the standardisation model, and thus biasing the findings. An alternative approach would have been to review student work to determine if the quality of performance was worthy of the CAG or standardised grade. The conclusion that the standardisation model limited low-probability grades to the extent that prevalence was much lower than in a normal year is a useful counterbalance to any argument of bias, however. Ultimately neither grading approach replicated all aspects of grading patterns seen in previous years, reflecting the trade-offs between key principles of assessment and definitions of standards set out in Section 3.

Section 5: General discussion and conclusions

The chance to perform: a golden principle

The UK's summer of 2020 was an examination series without examinations; of grades without the common summative assessments upon which they are usually based. A golden principle – that a student has some control over their grades by performing well in summative assessments – had to be violated. The extent to which students were able to outperform their centre's historical outcomes was linked to the extent to which the grading approach could account for their individual attainment. Four inputs were available for this: CAGs, centre rank orders, banked unit outcomes within the same qualification, and prior attainment data from previous qualifications.

As described in Section 1, the first grading process to publish results in 2020 was Scotland's, which unlike all others utilised CAGs directly within their model. In this respect, the model was therefore closest to the golden principle; its failure to garner public confidence set the public tone for the processes that followed. The *Banked Unit Model* did produce grades based on individuals' performance but was limited in use to A-levels in two UK nations, as well as a few GCSEs in Wales. The *Standard Model* of standardisation was used for linear qualifications where no unit outcomes for the qualification had been banked; exam boards were limited to using prior attainment data to account for changes in the strength of centres' respective cohorts but could not account for individuals' performance within their current course of study.

The rejection of all three models demonstrates the linkages between the UK's assessment systems. The effect of Scotland was to illustrate the flaws in the other nations' approaches, and ultimately none survived. Contagion had limits, however, demonstrating that this effect was not inevitable. For example, an algorithm was used to award grades in the Republic of Ireland, although it was amended following the UK experience by removing school prior attainment and historical cohort outcomes at the subject level (Kelly, 2021).

Three validity arguments, and their comparability and fairness logics

Whatever the differences, all standardisation models used in the UK ultimately represented an inadequate balance between the key principles of assessment because insufficient weighting was given to the one concurrent measure of attainment that was provided – the CAGs. The arguments against using CAGs as an absolute measure of attainment, such as inter-rater reliability and comparability in attainment standards between centres and across years, represent a particular fairness logic that was replicated from the UK's pre-pandemic assessment culture. CAGs could not meet a systemic definition of standards, applying only to the examinee level, yet CAGs represented the closest available to the requirement of the golden principle, and therefore the argument for their use won out.

The question remains, however: which argument? To respond to my first research question, as set out in Section 3, I identified two validity arguments made for CAGs, which differed with respect to the awarding phase. The first, building on an explanation of CAG grading patterns given by teachers in Ofqual's research (Holmes et al., 2021), was based on the notion that teachers systematically gave their students the benefit of the doubt when making grading decisions. The analysis presented in Section 4 gives further weight to that explanation. Except in Scotland, teachers were asked to predict grades, rather than to award them. Although the analysis in Section 4 focused on one qualification type in Wales, the finding that two or more grades would have represented a plausible outcome in a normal year is likely to apply elsewhere in the UK, given the similarities in qualification systems. With regard to my second research question, the analysis illustrates the difficulty of estimation at the student level. Selecting the higher grade more often than a cohort would normally achieve led to outcomes which, when aggregated across the cohort, resulted in implausible outcomes at that level. The policy goal of reconciling predicted grades at the student-level with comparable outcomes at the cohort level was doomed to fail, along with the standardisation models that operationalised it.

The incommensurability of student- and cohort-level considerations in 2020 should not be conflated, however, with the separate issue of CAGs being determined for students that significantly exceeded their prior attainment. These were the large 'downgrades' that featured in the Independent Review in Wales, and elsewhere – and while these made for a strong basis for press coverage, insufficient scrutiny was given to the cause of these discrepancies and volume of these cases. Conversely, my analysis showed that standardisation allowed for a smaller proportion of very unlikely (low-probability) grades to be awarded than would have been in a normal year – and the more constraints there were on students' opportunities to exceed expectations, the more the golden principle might have been seen to have been broken.

Trade-offs between comparability and validity occur wherever an assessment is repeated across settings, time periods and even between students (Sireci, 2021); this is reflected in the validity arguments identified in this study. While both validity arguments for CAGs are based on a re-prioritisation of principles away from comparability, the first is a temporary re-prioritisation that accounted for the need to place more weight on examinee-specific attainment data. The second validity argument for CAGs is a more fundamental departure from the usual balance of principles, stating that teachers are not so much more *optimistic* as more *knowledgeable* about a student's level of attainment, which the normal general qualifications assessment process also fails to capture, either because some students will have a bad day, or because outcomes are unfairly capped because of predictions used to ensure statistical comparability (Casella et al., 2021). Thus, the attainment measure from previous years was presented as insufficiently valid, and the notion of maintaining the attainment standard was rejected.

The evidence for achieving a high degree of reliability in teacher assessment is not strong. Johnson (2013) noted that this is not just true for individual pieces of work but also for portfolios of performances used to make overall holistic judgements on the level of attainment. Performance variability contributes to measurement error. Teachers' marking in high-stakes assessment is also

affected by the degree to which tasks and criteria are specified (Harlen, 2004) and knowledge of grade boundaries – for more subjectively-marked tasks, teachers may be inclined to mark upwards to a boundary (Isaacs, 2014). Even if the argument for validity is accepted, therefore, it may have come at the expense of reliability.

Interpretation of performances, and inferring attainment from them, is a key part of the validity arguments set out in the study. One of the main contributions of this study is to focus attention on authenticity as a core element of validity – both the nature of the assessment task and the conditions in which it is completed (Klenowski & Wyatt-Smith, 2010). These factors have long been known to have a significant impact on the performances that students produce and how we make inferences from them, but too often, debate centres on how we elicit performances and not on ensuring we do so in ways which allow us to make informed, appropriate inferences.

Awarding CAGs to prevent any under-grading also creates a bias, coming at the cost of allowing many students to achieve a higher grade than their likely level attainment, and thus relational fairness. In Section 4, my analysis demonstrated that grades with a low probability of being achieved based on AS performance were more common amongst CAGs than in a normal year. This highly risk-averse approach towards consequential and retrospective senses of fairness contradicts other forms of fairness (particularly relational fairness) and creates longer-term systemic risks, given the incentives it creates for teachers to over-grade to the benefit of their students. The awarding process has been adjusted previously to ensure consequential fairness to account for the sawtooth effect, but the scale of the adjustment needed to accept CAGs on this basis affected far more students.

There are broader concerns related to bias in teacher judgements. The final grades in 2020 led to wider attainment gaps relating to gender, free school meal eligibility, special educational needs, and broad ethnic group for the higher grades in general qualifications (Qualifications Wales, 2020h); however, the academic literature on bias in teacher assessment is more mixed (Lee & Walter,

2020). Although these gaps could not be attributed directly to bias amongst teachers, as the outcomes could be the result of the specific assessment model used or an alleged lack of training and support for teachers (Casella et al., 2021), students had expressed concern in 2020 that such biases could affect their outcomes (Qualifications Wales, 2020e).

The events of 2020 demonstrated the political nature of fairness: our understanding of the concept evolves, and can do so quickly. Given the value judgements required, however, this is not necessarily a bad thing. Often the same terminology was used to describe very different meanings for the key principles of assessment; this was particularly true of 'fairness', and my review highlights the need for 'fairness logics' (Baird, 2020) to be clearly set out with respect to standards, comparability, fairness and reliability. The UK's assessment culture's emphasis on comparability led to latent errors (Rhoades & Madaus, 2003); a miscalculation of how the fairness logic replicated from previous years and the usual balance of key principles would combine with weak data on student attainment and an (expected) optimism in teacher judgements in the eyes of the public. That included government, regulators and exam boards and stakeholders who, generally, offered qualified support for standardisation (Kippin & Cairney, 2021; Association of School and College Leaders, 2020).

Assessment cultures and contested space: 2020 as a catalyst for paradigm shifts?

My final research question was concerned with how 2020 may shape future developments in UK general qualifications. Latent error can be a trigger for paradigm shifts in assessment culture (Isaacs & Gorgen, 2018). The different validity arguments reflect not only positions on 2020 grading, but broader positions on how summative assessment and grading should take place. The Independent Reviews in Wales (Casella et al., 2021) and the Priestley review in Scotland (Priestley et al., 2020) linked the application of the models to the application of those principles in the pre-pandemic examination process. Both also linked to curriculum developments to propose a different balance of

principles and, over time, a different assessment culture. Subsequent reviews have led to conclusions which are striking in the similarity of their conclusions.

A recent review of Scotland's assessment system proposed a school graduation certificate incorporating school-based and external assessment as well as other contributions to the community, covering a broader range of attainments (Stobart, 2021). The International Educational Assessment Network (2020) – a group involving researchers involved in the development of Wales' new curriculum – made similar recommendations, proposing teacher-led continuous assessment, observation-based descriptions of personal capacities, less focus on fixed assessment at key age milestones, and increased provision of more useful information from national summative assessments to support their formative use. Wales' Independent Review echoed these views (Casella et al., 2021).

Underpinning this is the principle that the assessment process – and standard-setting processes more specifically – needs to reflect children's rights and the need for the education system to act in their best interest, and that the focus of how we conceptualise fairness needs to move from the group level (cohort, centre, demographic group) to a more radical concept that focuses on each student's assessment needs, giving them every opportunity to perform (Woods et al., 2019). The focus on observed performance and a student-level definition of standards reflect both the second validity argument for CAGs – teachers' knowledge of students enabling them to make more valid and useful judgements on student's attainment on an individual basis – and the outcomes-based assessment paradigm, rather than the curriculum-based assessment paradigm within which the current general qualifications are situated.

Baird and Grey (2016) noted that what is acceptable or controversial in assessment matters is a matter of culture. The Covid-19 pandemic has opened spaces for discussing what the four nations of the UK need from their general qualifications systems, as elsewhere, but it is not yet clear if this will lead to lasting or profound change in the UK assessment system. Isaacs and Gorgen framed Kuhn's

(1962) work to identify three preconditions for paradigm change. The first, dissatisfaction with the current accepted paradigm, existed before 2020. Pre-existing critiques of the examination system, such as the use of statistics in awarding preventing a 'forgotten third' of students from passing key GCSEs (Association of School and College Leaders, 2019) gained new power once standardisation algorithms had been rejected. The second, agreement on an alternative as a better fit, seems closer in Scotland and Wales because of curriculum developments, which require changes in qualifications (Stobart, 2021; Qualifications Wales, 2021). While this study has found that there are advocates for a different approach to assessment, the question of whether the third precondition is met, that they outnumber or outweigh those for the old paradigm, is currently unclear. This last precondition is therefore likely to be pivotal in determining if 2020 is to have a longer-term impact.

Assessment is an inferential process. A student-centred approach to standard-setting requires a nuanced and careful professional framework for making those inferences, a strong understanding of the content standard being applied, the nature of assessment tasks and conditions, and the effect these factors have on student performance. As the events of 2020 show, all key principles of assessment matter, even the ones we take for granted.

References

AERA, APA and NCME (2014). *The Standards for Educational and Psychological Testing, 2014 edition.* Washington, D.C.: American Educational Research Association (AERA).

Anders, J., Dilnot, C., Macmillian, L. and Wyness, G. (2020). Grade Expectations: How Well Can We Predict Future Grades Based On Past Performance? No 20-14, CEPEO Working Paper Series. London: Centre for Education Policy and Equalising Opportunities, UCL Institute of Education.

AQA (2021). *Quality Assessment*. Accessed 30th January 2021 from https://www.aqa.org.uk/about-us/what-we-do/getting-the-right-result/quality-assessment

Association of School and College Leaders (2019). *The Forgotten Third: Final Report of the Commission of Inquiry.* Leicester: Association of School and College Leaders.

Association of School and College Leaders (2020). ASCL Cymru Comments on Qualifications Wales Announcement. Accessed 11th August 2021 from https://www.ascl.org.uk/News/Our-news-and-press-releases/ASCL-Cymru-comments-on-Qualifications-Wales-announ

Atkinson, P. (2015). *For Ethnography*. London: SAGE Publications Ltd.

Baird, J. (1997). *Teachers' Estimates of A-level Performance.* Guildford: Associated Examining Board (internal report).

Baird, J. (2018a). *On different paradigms of educational assessment. Implications for the relationships with teaching and learning*. 37th Vernon-Wall Lecture, 27th October 2017. Leicester: British Psychological Society.

Baird, J. (2018b). The Meaning of National Examination Standards. In: Baird, J., Isaacs, T., Opposs, D. and Gray, L. *Examination Standards: How Measures and Meanings Differ Around the World* (pp. 284-306). London: UCL Institute of Education Press.

Baird, J. (2020). What are the Implications of 2020 Grade Inflation? *Schools Week*, August 21st. Accessed 15th August from https://www.tes.com/news/what-are-implications-2020-grade-inflation

Baird, J. (2021). We Need to Talk About Our Assessment System. *Schools Week*, June 11th. Accessed 11th August 2021 from https://schoolsweek.co.uk/we-need-a-national-conversation-about-assessment/

Baird, J., Caro, D., Elliott, V., El Masri, Y., Ingram, J., Isaacs, T., Pinot de Moira, A., Randhawa, A., Stobart, S., Meadows, M., Morin, C. and Taylor, R. (2019). *Examination Reform: Impact of Linear and Modular Examinations at GCSE.* Oxford: Oxford University Centre for Educational Assessment; Coventry: Ofqual.

Baird, J. and Gray, L. (2016). The Meaning of Curriculum-Related Examination Standards in Scotland and England: A Home-International Comparison. *Oxford Review of Education*, 42:3, pp.266-284.

BBC (2020a). *Coronavirus in Scotland: School Pupils Protest Over SQA Exam Results, 7th August 2020.* Accessed 3rd January 2021 from https://www.bbc.co.uk/news/uk-scotland-53692135

BBC (2020b). *GCSEs and A-level Exams: Teacher Estimates to Be Used for Welsh Grades.* Accessed 12th August from https://www.bbc.co.uk/news/uk-wales-53807854

BBC (2021). *Northern Ireland cancels GCSE and A-level exams.* Accessed 31st July 2021 from https://www.bbc.co.uk/news/uk-northern-ireland-55560314

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Zand Scholten, A. and S. Franić (2009). The End of Construct Validity. In: Lissitz, R. W. *The Concept of Validity: Revisions, New Directions, and Applications* (pp.39-64). Charlotte, NC: Information Age Publishing.

British Educational Research Association (2018). *Ethical Guidelines for Educational Research*, 4th edition. London: British Educational Research Association.

Bryman, A. (2012). *Social Research Methods*, 4th edition. Oxford: Oxford University Press.

Casella, L., Harris, A., Youell, A. and Jones, R. (2020). *Independent Review of the Summer 2020 Arrangements to Award Grades, and Considerations for Summer 2021: Interim Report, October 2020.* Retrieved 17th November 2020 from https://gov.wales/sites/default/files/publications/2020-10/independent-review-of-the-summer-2020-arrangements-to-award-grades-and-considerations-for-summer-2021-interim-report-october-2020.pdf

Casella, L., Harris, A., Youell, A. and Jones, R. (2021). *Independent Review of the Summer 2020 Arrangements to Award Grades, and Considerations for Summer 2021: Final Report.* Accessed 22nd January 2021 from https://gov.wales/sites/default/files/publications/2021-01/independent-review-of-the-summer-2020-arrangements-to-award-grades-and-considerations-for-summer-2021.pdf

CCEA (2020a). *Council*. Retrieved 27th December 2019 from https://ccea.org.uk/about/council

CCEA (2020b). *Information on GCE and GCSE*. Retrieved 27th December 2019 from https://ccea.org.uk/regulation/qualifications-regulation/information-gce-and-gcse

CCEA (2020c). *Compendium of Approaches Taken and Formulae Used in GCE and GCSE Calculation of Grades, Summer 2020*. Retrieved 3rd January 2021 from https://ccea.org.uk/downloads/docs/ccea-asset/General/Compendium%20of%20approaches%20taken%20and%20formulae%20used%20in%20GCE%20and%20GCSE%20calculation%20of%20grades%2C%20Summer%202020.pdf

CCEA (2020d). *Statement by CCEA in Relation to Summer 2020 Awarding for A / AS levels*. Retrieved 3rd January 2021 from https://ccea.org.uk/news/2020/august/statement-ccea-relation-summer-2020-awarding-levels

Children, Young People and Education Committee (2020a). *Record of Committee, 5th November*. Accessed 16th August 2021 from https://record.assembly.wales/Committee/6545

Children, Young People and Education Committee (2020b). *Record of Committee, 18th August*. Accessed 16th August from https://record.assembly.wales/Committee/6452

Coe, R. (2010). Understanding Comparability of Examination Standards. *Research Papers in Education*, 25(3), pp.271–284.

Costley, C., Elliott, G. and Gibbs, P. (2010). *Doing Work-Based Research: Approaches to Enquiry for Insider-Researchers*. London: SAGE Publications Ltd.

Cresswell, M.J. (2003). *Heaps, Prototypes and Ethics: The Consequences of Using Judgements of Student Performance to Set Examination Standards in a Time of Change*. London, UK: University of London Institute of Education.

Crisp, V. and Shaw, S.D. (2012). Applying Methods to Evaluate Construct Validity in the Context of A-level Assessment. *Educational Studies*, 38:2, pp.209-222.

Crooks, T. J., Kane, M.T. and Cohen, A.S. (1996). Threats to the Valid Use of Assessments, *Assessment in Education: Principles, Policy & Practice*, 3:3, pp.265-286.

Cuff, B.M.P., Meadows, M. and Black, B. (2019). An Investigation into the Sawtooth Effect in Secondary School Assessments in England. Assessment in Education: Principles, Policy & Practice, 26:3, pp.321-339.

Department for Education (2020a). *Schools, Colleges and Early Years Settings to Close*. Retrieved 27th December 2020 from https://www.gov.uk/government/news/schools-colleges-and-early-years-settings-to-close

Department for Education (2020b). *Direction under S129(6) of the Apprenticeships, Skills, Children and Learning Act 2009, 31st March 2020*. Retrieved 3rd January 2021 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/877611/Letter_from_Secretary_of_State_for_Education_to_Sally_Collier.pdf

Department for Education (2020c). *Letter from Gavin Williamson MP, Secretary of State for Education, to Dame Glenys Stacey, Interim Chief Regulator, Ofqual, 12th October 2020.* Retrieved 17th November 2020 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/926239/letter_from_DfE_Gavin_Williamson_to_Ofqual_Glenys_Stacey.pdf

Department for Education (2020d). *Coronavirus (COVID-19): School and College Performance Measures*. Accessed 31st July 2021 from https://www.gov.uk/government/publications/coronavirus-covid-19-school-and-college-performance-measures/coronavirus-covid-19-school-and-college-accountability-for-the-2019-to-2020-academic-year

Department for Education and Ofqual (2021). Consultation Launched on 2021 Exam Replacement. Accessed 31st July 2021 from https://www.gov.uk/government/news/consultation-launched-on-2021-exam-replacement

Dhillon, D. (2005). Teachers' Estimates of Candidates' Grades: Curriculum 2000 Advanced Level Qualifications. *British Educational Research Journal*, 31:1, pp.69-88.

Donaldson, G. (2015). *Successful Futures: Independent Review of Curriculum and Assessment Arrangements in Wales*. Accessed 30th August 2021 from https://gov.wales/sites/default/files/publications/2018-03/successful-futures.pdf

Elliott, G. (2011). A Guide to Comparability Terminology and Methods. *Research Matters: A Cambridge Assessment Publication*, Special Issue 2, pp.9-19.

Flockhart, G. (2020). *Coronavirus in Scotland RECAP: Pupils to Receive Original Marks Estimated by Teachers, 11th August 2020*. Retrieved 3rd January 2021 from https://www.scotsman.com/health/coronavirus/coronavirus-scotland-recap-pupils-receive-original-marks-estimated-teachers-2938629

Furr, R. M. (2018). *Psychometrics*, 3rd Edition. Thousand Oaks, CA: SAGE Publications.

Gipps, C.V. and Murphy, P. (1994). *A Fair Test? Assessment, Achievement and Equity.* Buckingham: Open University Press.

Gray, L. (2018). Research in National Examination Standards as an Insider. In: Baird, J-A, Isaacs, T., Opposs, D. and Gray, L. (eds.), *Examination Standards: How Measures and Meanings Differ Around the World* (pp.41-53). London: UCL Institute of Education Press.

Harlen, W. (2004). A Systematic Review of the Evidence of Reliability and Validity of Assessment by Teachers Used for Summative Purposes. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Harlen, W. (2005) Trusting Teachers' Judgement: Research Evidence of the Reliability and Validity of Teachers' Assessment Used for Summative Purposes, *Research Papers in Education*, 20:3, pp.245-270.

Harrell, Jr., F.E. (2015). *Regression Modelling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd edition. Switzerland: Springer International Publishing.

Harris, A. [@AlmaHarris1] (2021, 6 August). What if grade deflation has been the norm over many, many years, using statistical models rather than professional judgement, and this is simply recalibration. [Tweet]. Twitter. https://twitter.com/AlmaHarris1/status/1423594390484828160

He, Q., Opposs, D., Glanville, M. and Lampreia-Carvalho, F. (2015). Assessing Pupils at the Age of 16 in England – Approaches for Effective Examinations. *The Curriculum Journal*, 26(1), pp.70-90.

Holmes, S., Churchward, D., Howard, E., Keys, E., Leahy, F., Tonin, D. and Black, B. (2021). *Centre Judgements: Teaching Staff Interviews, Summer 2020*. Coventry: Ofqual.

House of Commons Education Committee (2020). *Letter from Robert Halfon MP, Chair to the Secretary of State for Education, 10th November 2020: Cancellation of 2020 Exams and 2021 Exams Series.* Retrieved 17th November 2020 from https://committees.parliament.uk/publications/3401/documents/32495/default/

Information Commissioner's Office (2020). Statement in Response to Exam Results, 14th August 2020. Retrieved 17th November 2020 from https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2020/08/statement-in-response-to-exam-results/

International Educational Assessment Network (2020). *Policy Conversation Paper: Learning Progression: Implications for Curriculum & Assessment.* Glasgow: International Educational Assessment Network.

Isaacs, T. (2014). Curriculum and Assessment Reform Gone Wrong: The Perfect Storm of GCSE English, *Curriculum Journal*, 25:1, pp.130-147.

Isaacs, T. and Gorgen, K. (2018). Culture, Context and Controversy in Setting National Examination Standards. In: Baird, J., Isaacs, T., Opposs, D. and Gray, L. *Examination Standards: How Measures and Meanings Differ Around the World* (pp.307-330). London: UCL Institute of Education Press.

JCQ (2020a). *Instructions for Conducting Examinations: 1st September 2020 to 31 August 2021.* London: Joint Council for Qualifications.

JCQ (2020b). *Access Arrangements, Reasonable Adjustments and Special Consideration.* Accessed 15th September 2021 from https://www.jcq.org.uk/exams-office/access-arrangements-and-special-consideration/#:~:text=Home%20%C2%BB%20Exams%20Office%20%C2%BB%20Access%20Arrangements%2C%20Reasonable,evidence%20of%20need%20and%20normal%20way%20of%20working.

JCQ (2020c). *A Guide to the Special Consideration Process – General and Vocational Qualifications – With effect from 1 September 2020*. Accessed 26th January 2021 from https://www.jcq.org.uk/wp-content/uploads/2020/08/A-guide-to-the-spec-con-process-202021-Website-version.pdf

JCQ (2020d). *Supplementary Information on Validation of Evidence by Centres*. Accessed 11th August from https://www.wjec.co.uk/media/qxridh2n/supplementary-information-on-validation-of-evidence-by-centres.pdf

Johnson, R.B., Onwuegbuzie, A.J. and Turner, L.A. (2007). Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research*, 1:2 (April), pp.112-133.

Johnson, S. (2013). On the Reliability of High-Stakes Teacher Assessment, *Research Papers in Education*, 28:1, pp.91-105.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50:1 (Spring), pp.1-73.

Kelly, A. (2021). A Tale of Two Algorithms: The Appeal and Repeal of Calculated Grades Systems in England and Ireland in 2020. *British Educational Research Journal*, 47:3, pp.725–741.

Kippin, S. and Cairney, P. (2021). The COVID-19 Exams Fiasco across the UK: Four Nations and Two Windows of Opportunity. *British Politics*, https://doi.org/10.1057/s41293-021-00162-y

Klenowski, V., and Wyatt-Smith, C.M. (2010). Standards, Teacher Judgement and Moderation in Contexts of National Curriculum and Assessment Reform. *Assessment Matters*, 2, pp.107–131.

Kuhn, T.S. (1962). The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Lee, M-W and Walter, M. (2020). *Equality Impact Assessment: Literature Review*. Coventry: Ofqual.

Long, J.S. and Freese, J. (2014). *Regression Models for Categorical Dependent Variables Using Stata*, 3rd Edition. College Station, TX: Stata Press.

Martin, J. (2018). Historical and Documentary Research, pp.323-333. In: Cohen, L., Manion, L. and Morrison, K., *Research Methods in Education*, 8th edition. Abingdon, UK and New York: Routledge.

Matchett, C. (2020). Top students 'pulled down' as academic labels Scottish exam results moderation 'unfair' and 'arbitrary'. The Scotsman, 5th August. Accessed 15th September 2021 from https://www.scotsman.com/education/top-students-pulled-down-academic-labels-scottish-exam-results-moderation-unfair-and-arbitrary-2933120

Maul, A. (2013). On the Ontology of Psychological Attributes. *Theory & Psychology*, 23:6, pp.752–769.

Maxwell, J.A. (2017). The Validity and Reliability of Research: A Realist Perspective, pp.116-141. In: Wyse, D., Selwyn, N., Smith, E. and Suter, L.E., *The BERA/SAGE Handbook of Educational Research*. London: SAGE Publications Ltd.

McFadden, D. (1979). Quantitative Methods for Analysing Travel Behaviour on Individuals: Some Recent Developments. pp.279-313 in: Hensher, D. and Stopher, P., (eds.). *Behavioural Travel Modelling: International Conference on Behavioural Travel Modelling, 1977, Tanunda, South Australia.* London: Croom Helm.

Medical Schools Council (2021). *Briefing Paper: Confirmed Places for Undergraduate Medicine Courses on A-level Results Day 2021*. Accessed 12th August 2021 from https://www.medschools.ac.uk/media/2883/briefing-paper-a-level-results-day-2021.pdf

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, 50:9, pp.741-749.

Moore, R. (2013). Social Realism and the Problem of the Problem of Knowledge in the Sociology of Education. *British Journal of Sociology of Education*, 34:3, pp.333-353.

NAHT Cymru (2020). *NAHT Cymru Statement in Response to A-level Results Protests in Wales*. Accessed 11th August from https://www.naht.org.uk/About-Us/NAHT-Cymru/ArtMID/606/ArticleID/714/NAHT-Cymru-statement-in-response-to-A-level-results-protests-in-Wales

Newton, P.E. (2010a). Contrasting Conceptions of Comparability. *Research Papers in Education*, 25:3, pp.285-292.

Newton, P.E. (2010b). Thinking About Linking. *Measurement: Interdisciplinary Research and Perspectives*, 8:1, pp.38-56.

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

Newton, P.E. (2013). Two Kinds of Argument? *Journal of Educational Measurement*, 50:1, pp.105-109.

Newton, P.E. (2017). There Is More to Educational Measurement Than Measuring: The Importance of Embracing Purpose Pluralism. *Educational Measurement: Issues and Practice*, 36:2, pp.5-15.

Newton, P.E. (2020). *Maintaining Standards: During Normal Times and When Qualifications are Reformed*. Coventry: Ofqual.

Newton, P.E. and Baird, J-A. (2016). The Great Validity Debate. *Assessment in Education: Principles, Policy & Practice*, 23:2, pp.173-177.

Newton, P.E. and Shaw, S.D. (2014). *Validity in Educational & Psychological Assessment*. London: SAG Publications; Cambridge, UK: Cambridge Assessment.

Nisbet, I. and Shaw, S. (2020). *Is Assessment Fair?* London: SAGE Publications Ltd.

Northern Ireland Department for Education (2020). *Letter from the Office of the Minister of Education to All School Principals, regarding Delivery of GCSE, AS and A-level Qualifications in 2020/21, 9th October 2020*. Retrieved 17th November 2020 from https://www.education-ni.gov.uk/sites/default/files/publications/education/Letter%20to%20Principals_0.pdf

OECD (2021). *Scotland's Curriculum for Excellence: Into the Future.* Accessed 29th August from https://www.oecd.org/education/scotland-s-curriculum-for-excellence-bf624417-en.htm

Office for Statistical Regulation (2020). OSR Review of Approach to Developing Statistical Models Designed for Awarding 2020 Exam Results. Retrieved 17th November 2020 from https://osr.statisticsauthority.gov.uk/our-regulatory-work/osr-review-of-approach-to-developing-statistical-models-designed-for-awarding-2020-exam-results/

Ofqual (2019a). *Ofqual Corporate Plan 2019-2022.* Retrieved 27th December 2020 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/800458/Ofqual_Corporate_Plan_2019_-_22.pdf

Ofqual (2019b). *Ofqual Governance Framework, revised November 2019*. Retrieved 27th December 2019 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/864633/Ofqual_Governance_Framework_Nov_2019.pdf

Ofqual (2019c). *Summer 2019 Data Exchange Procedures GCE (AS/A-level), GCSE and Project Qualifications.* Coventry: Ofqual.

Ofqual (2020a). *Provisional Entries for GCSE, AS and A-level: Summer 2020 Exam Series*. Retrieved 27th December 2020 from https://www.gov.uk/government/publications/provisional-entries-for-gcse-as-and-a-level-summer-2020-exam-series/provisional-entries-for-gcse-as-and-a-level-summer-2020-exam-series

Ofqual (2020b). *Regulating GCSEs, AS and A-levels: Guide for Schools and Colleges.* Retrieved 27th December 2020 from https://www.gov.uk/guidance/regulating-gcses-as-and-a-levels-guide-for-schools-and-colleges/about-gcses-as-and-a-levels

Ofqual (2020c). *Process for Awarding GCSEs, AS and A-levels Summer 2020*. Retrieved 3rd January 2021 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/887041/6614-4_Infographic_-_Process_for_awarding_GCSEs__AS_and_A_levels_summer_2020.pdf

Ofqual (2020d). *Consultation Decisions: Exceptional Arrangements for Exam Grading and Assessment in 2020: Consultation on Specified General Qualifications – GCSEs, AS, A-levels, Extended Project Qualifications and the Advanced Extension Award in Maths.* Retrieved 3rd January 2021 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/887048/Summer_2020_Awarding_GCSEs_A_levels_-_Consultation_decisions_22MAY2020.pdf

Ofqual (2020e). *Statement from Roger Taylor, Chair, Ofqual*. Retrieved 3rd January 2021 from https://www.gov.uk/government/news/statement-from-roger-taylor-chair-ofqual

Ofqual (2020f). *Awarding GCSE, AS, A-level, Advanced Extension Awards and Extended Project Qualifications in Summer 2020: Interim Report*. Accessed 22nd February 2021 from 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

Ofqual (2021). *Summary Statistics for GCSE, AS and A-level*. Accessed 28th August 2021 from https://analytics.ofqual.gov.uk/apps/GCSEandGCE/SummaryStats/

Ofqual, Qualifications Wales and CCEA Regulator (2019a). *Statement from the Qualification Regulators on Changes to GCSEs, AS and A-levels*. Retrieved 27th December 2020 from https://www.qualificationswales.org/media/4950/statement-from-the-qualification-regulators-eng.pdf

Ofqual, Qualifications Wales and CCEA Regulator (2019b). *AS and A-levels in England, Wales and Northern Ireland: Infographic*. Retrieved 27th December 2020 from https://ccea.org.uk/downloads/docs/regulation-asset/Qualifications%20Regulation/Common%20features%20and%20key%20differences%20in%20qualifications%20in%20England%2C%20Wales%20and%20Northern%20Ireland%20-%20Infographic.pdf

Opposs, D. and Gorgen, K. (2018). What is Standard Setting? In: Baird, J., Isaacs, T., Opposs, D. and Gray, L. *Examination Standards: How Measures and Meanings Differ Around the World* (pp.54-76). London: UCL Institute of Education Press.

Pearson (2021). *Moderation and Internal Assessment for General Qualifications*. Accessed 31st January 2021 from https://qualifications.pearson.com/en/support/support-topics/assessment-and-verification/moderation-and-internal-assessment.html

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (2001). *Knowing What Students Know. The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.

Plaid Cymru (2020). *Free and Independent Appeals Process is Vital for A-level Results Fairness.* Accessed 12th August 2021 from https://www.partyof.wales/a_level_results_fairness

Priestley, M., Shapira, M., Priestley, A., Ritchie, M and C. Barnett (2020). *Rapid Review of National Qualifications Experience 2020*. Edinburgh: The Scottish Government. Retrieved 17th November 2020 from 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

Prior, L. (2008). Repositioning Documents in Social Research. *Sociology*, Special Issue on Research Methods, 42, pp.821-836.

Prior, L. (2016). Using Documents in Social Research, pp.171-185. In: Silverman, D. (ed.), *Qualitative Research*, 4th edition. London: SAGE Publications Ltd.

Qualifications Wales (2017). *General Qualifications Strategic Plan, July 2017*. Retrieved 27th December 2020 from https://www.qualificationswales.org/english/publications/gq-strategy-2017/

Qualifications Wales (2019a). *Wales Summer 2019 Data Exchange Procedures.* Accessed 22nd February 2021 from https://www.qualificationswales.org/english/publications/wales-summer-2019-data-exchange/

Qualifications Wales (2019b). *Variation in GCSE, AS and A-level Qualification Results in Wales, Summer 2020*. Accessed 8th August 2021 from https://qualificationswales.org/media/5172/release.pdf

Qualifications Wales (2020a). *Provisional Entries for Summer 2020 GCSEs, AS, A-levels, Skills Challenge Certificate, and Level 1/Level 2 Certificates for Wales*. Retrieved 27th December 2020 from https://www.qualificationswales.org/media/5998/summer-entries-2020-release.pdf

Qualifications Wales (2020b). *Corporate Governance: Board*. Retrieved 27th December 2019 from https://www.qualificationswales.org/english/about-us/corporate-governance/bwrdd/

Qualifications Wales (2020c). *Designation of Qualifications*. Retrieved 27th December 2019 from https://www.qualificationswales.org/english/our-work/regulating-qualifications/designation-of-qualifications/

Qualifications Wales (2020d). *Awarding in 2021: FAQs*. Retrieved 28th December 2020 from https://qualificationswales.org/english/covid-19-awarding-2021/faqs/

Qualifications Wales (2020e). *Arrangements for Summer 2020 Exam Series: Decisions Taken Following Consultation*. Retrieved 3rd January 2021 from https://qualificationswales.org/media/6181/arrangements-for-summer-2020-exam-series-decisions-taken-following-consultation.pdf

Qualifications Wales (2020f). *Welsh Government Direction to Qualifications Wales*. Retrieved 3rd January 2021 from https://www.qualificationswales.org/english/covid-19---results-2020/summer-2020/welsh-government-statements/welsh-government-direction-to-qualifications-wales/

Qualifications Wales (2020g). *Protecting the Integrity of Qualifications in Wales.* Accessed 11th August 2021 from https://qualificationswales.org/english/news/protecting-the-integrity-of-qualifications-in-wales/

Qualifications Wales (2020h). *Summer 2020 Equalities Impact Analysis: GCSE, AS, and A-level.* Accessed 11th August 2021 from https://qualificationswales.org/media/6744/equalities-impact-analysis-publication.pdf

Qualifications Wales (2021). *Qualified for the Future: The Right Choice for Wales.* Accessed 26th June 2021 from: https://qualificationswales.org/media/6997/qualified-for-the-future-the-right-choice-for-wales-consultation.pdf.

Raffe, D. (2006). Devolution and Divergence in Education Policy, pp.52-69. In: Adams, J. and Schmuecker, K., *Devolution in Practice*. Newcastle: IPPR North.

Raffe, D., Brannen K., Croxford, L. and Martin, C. (1999). Comparing England, Scotland, Wales and Northern Ireland: The Case for 'Home Internationals' in Comparative Research. *Comparative Education*, 35:1, pp.9-25.

Rhead, S., Black, B. and Pinot de Moira, A. (2018). *Marking Consistency Metrics: An Update*. Accessed 10th February 2021 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/759207/Marking_consistency_metrics_-_an_update_-_FINAL64492.pdf

Rhead, S., He, Q., Black, B. and Patchett, D. (2018). *What Causes Variability in Centre Level GCSE Results Year-on-Year? Some further analysis*. Coventry: Ofqual.

Rhoades, K. and Madaus, G. (2003). Errors in Standardized Tests: A Systemic Problem. Boston, MA: Boston College/National Board on Educational Testing and Public Policy. Accessed 13th September 2021 from https://www.bc.edu/research/nbetpp/statements/M1N4.pdf

Richmond, T. (2021). *Re-assessing the Future: Part 1 – How to Move Beyond GCSEs.* London: EDSK.

Scott, J. (1990). *A Matter of Record: Documentary Sources in Social Research*. Cambridge: Polity Press.

Scottish Government (2020a). *Deputy First Minister - SQA 2020 Results: Statement given by Education Secretary John Swinney to the Scottish Parliament on 11 August 2020.* Retrieved 2nd January 2021 from https://www.gov.scot/publications/deputy-first-minister-sqa-2020-results/

Scottish Government (2020b). *National Qualifications Experience 2020 - Rapid Review: Our Response*. Retrieved 17th November 2020 from https://www.gov.scot/publications/rapid-review-of-national-qualifications-experience-2020-our-response/

Scottish Government (2020c). *SQA Awards Update*. Accessed 1st January 2021 from https://www.gov.scot/news/sqa-awards-update/

Scottish Government (2021). *OECD Review Backs School Curriculum*. Accessed 29th August 2021 from https://www.gov.scot/news/oecd-review-backs-school-curriculum/

Sherwood, D. (2020). *Why 'Exams as Usual' Are a Bad Idea*. Published 15th September. Accessed 30th January 2021 from https://www.hepi.ac.uk/2020/09/15/dennis-sherwood-why-exams-as-usual-are-a-bad-idea/

Sireci, S. (2021). Standardization and UNDERSTANDardization in Educational Assessment. *Educational Measurement: Issues and Practice*, 39:3, pp.100-105.

Smith, M. (2020). *How likely is likely? YouGov explores people's perceptions of probabilities.* Accessed 1st October 2020 from www.yougov.co.uk/topics/science/articles-reports/2020/10/01/how-likely-likely

SQA (2014). *The New National Qualifications — New Opportunities for Your Child: A Guide for Parents and Carers*. Retrieved 28th December 2020 from https://www.education.gov.scot/parentzone/Documents/SQACfEParentsGuideJun14.pdf

SQA (2019). *Annual Statistical Report 2019*. Retrieved 28th December 2020 from https://www.sqa.org.uk/sqa/91419.html

SQA (2020a). *Who We Are.* Retrieved 28th December 2020 from https://www.sqa.org.uk/sqa/79159.2325.html

SQA (2020b). *Technical Report: National Qualifications 2020 Awarding – Methodology Report*. Retrieved 3rd January 2021 from https://www.sqa.org.uk/sqa/files_ccc/SQAAwardingMethodology2020Report.pdf

SQA (2020c). *Moderation Summary Report Subject: English Level: Higher*. Retrieved 3rd January 2021 from https://www.sqa.org.uk/sqa/files_ccc/H-estimate-moderation-summary-English.pdf

SQA (2020d). *2020 Alternative Certification Model: Equality Impact Assessment*. Retrieved 3rd January 2021 from https://www.sqa.org.uk/sqa/files_ccc/2020-sqa-alternative-certification-model-equality-impact-assessment.pdf

StataCorp (2021). Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC.

Stenhouse, L. (1975). *An Introduction to Curriculum Research and Development.* London: Heinemann Educational.

Stobart, G. (2021). Upper-Secondary Education Student Assessment in Scotland: A Comparative Perspective. OECD Education Working Paper No. 253. Accessed 12th September 2021 from https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/WKP(2021)10&docLanguage=En

Taylor, R. and Opposs, D. (2018). Standard Setting in England: A-levels. In: Baird, J., Isaacs, T., Opposs, D. and Gray, L. *Examination Standards: How Measures and Meanings Differ Around the World* (pp.100-113). London: UCL Institute of Education Press.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge, UK: Cambridge University Press.

UCAS (2016). *Factors Associated with Predicted and Achieved A-level Attainment*. Accessed 31st July from https://www.ucas.com/file/71796

Webb, E.J., Campbell, D.T., Schwartz, R.D. and Sechrest, L. (1966). *Unobtrusive Measures: Nonreactive Methods in the Social Sciences.* Chicago: Rand McNally.

Weinberg, J. (2020). *Ofqual and Data*-sharing. Accessed 26th January 2021 from https://ofqual.blog.gov.uk/2020/12/21/ofqual-and-data-sharing/

Welsh Government (2020a). *Education Minister Announces Exam Arrangements for Years 10 and 12*. Accessed 31st July from https://gov.wales/education-minister-announces-exam-arrangements-years-10-and-12

Welsh Government (2020b). *Independent Review of the Summer 2020 Arrangements to Award Grades, and Considerations for Summer 2021: Terms of Reference.* Retrieved 17th November 2020 from https://gov.wales/sites/default/files/publications/2020-09/independent-review-summer-2020-arrangements-award-grades-considerations-summer-2021_0.pdf

Welsh Government (2020c). *Wales' Approach for Qualifications in 2021 Confirmed by Education Minister Kirsty Williams, 10th November 2020.* Retrieved 17th November 2020 from https://gov.wales/wales-approach-qualifications-2021-confirmed-education-minister-kirsty-williams

Welsh Government (2020d). *Direction for the Summer 2020 Series.* Retrieved 3rd January 2021 from https://qualificationswales.org/media/5850/gcse-as-and-a-level-cancellations-letter.pdf

Welsh Government (2020e). *Guidance: School Performance Reporting Arrangements: Important Update*. Accessed 31st July from https://gov.wales/school-performance-reporting-arrangements-important-update-html

Winkley, J. and Cresswell, M. (2012). An Introduction to the Concept of Reliability in the Context of Educational Assessment, pp.3-8. In: Opposs, D. and He, Q. (eds.), *Ofqual's Reliability Compendium*. Coventry: Office of Qualifications and Examinations Regulation.

WJEC (2020a). *Methods Report: Qualifications Wales-regulated A-levels*. Cardiff: WJEC-CBAC Ltd.

WJEC (2020b). *Methods Report: Qualifications Wales-regulated AS*. Cardiff: WJEC-CBAC Ltd.

WJEC (2020c). *WJEC GCSE Physical Education: Teacher Guidance – Calculating Centre Assessment Grades in Summer 2020*. Cardiff: WJEC-CBAC Ltd.

WJEC and City & Guilds (2020). *Health and Care Learning Wales: The Suite of Qualifications.* Accessed 27th December 2020 from http://wjec-website.azurewebsites.net/media/1984/wjec-vs-city-guilds_qualifications-offer_jan20_english.pdf

Wolf, A. (2001). Competence-Based Assessment. In: Raven, J. and Stephenson, J. (eds)., *Competence in the Learning Society* (pp.453-466). New York, NY: Peter Lang Publishing, Inc.

Woods, K., McCaldin, T., Hipkiss, A., Tyrrell, B. and Dawes, M. (2019). Linking Rights, Needs, and Fairness in High-Stakes Assessments and Examinations, *Oxford Review of Education*, 45:1, pp.86-101.

Yen, C-J and Liu, S. (2009). Learner Autonomy as a Predictor of Course Success and Final Grades in Community College Online Courses, *Journal of Educational Computing Research*, 41:3, pp.347-367.

Yin, R.K. (2018). *Case Study Research and Applications: Design and Methods*, sixth edition. Thousand Oaks, CA: SAGE Publications Ltd.

Young, M. (2008). From Constructivism to Realism in the Sociology of the Curriculum. *Review of Research in Education*, 32, pp.1-28.

Appendix 1: CUREC application

CENTRAL UNIVERSITY
RESEARCH ETHICS COMMITTEE (CUREC)

CUREC 1A Checklist
for the Social Sciences and Humanities

The University of Oxford places a high value on the knowledge, expertise, and integrity of its members and their ability to conduct research to high standards of scholarship and ethics. The research ethics clearance procedures have been established to ensure the University is meeting its obligations as a responsible institution. They start from the presumption that all members of the University take their responsibilities and obligations seriously and will ensure that their research involving human participants is conducted according to the established principles and good practice in their fields and in accordance, where appropriate, with legal requirements. Since the requirements of research ethics review will vary from field to field and from project to project, the University accepts that different guidelines and procedures will be appropriate.

Please refer to Where and how to apply for ethical review and the CUREC flowchart first to see if you need ethics approval.

Please complete this form using a word processor and email it, together with your supporting documents, to your Departmental Research Ethics Committee (DREC) (if applicable). If you don't have a DREC please email this form to ethics@socsci.ox.ac.uk using your official ox.ac.uk email address. Only type-written, emailed applications will be accepted.

SECTION A: Filter for CUREC 2 application		
This section determines whether your study raises more complex issues requiring the completion of a full application for ethical review, known as the CUREC 2 application. **(Please mark 'X' in the Yes/ No column.)**		
1. Are research participants classed as people whose ability to give free and informed consent is in question? (This may include under 18s (although see competent youths), prisoners, or adults at risk.) Your attention is drawn to the University's Safeguarding Code of Practice and its implications for researchers involving children or adults at risk. This includes the need for the work to be risk assessed and for researchers to undertake related training. (**Note**: If any of your participants are aged 16 or under, answer 'Yes' here **and** also answer question 5 below.)	Yes ☐	No ☒
2. By taking part in the research, will participants be at risk of criminal prosecution (e.g. by providing information on drug abuse or child abuse)?	Yes ☐	No ☒
3. Does the research involve the deception of participants?	Yes ☐	No ☒
4. Does your research raise issues relevant to the Counter-Terrorism and Security Act (the Prevent duty), which seeks to prevent people from being drawn into terrorism? Please see advice on this on our Best Practice Guidance web page.	Yes ☐	No ☒
If you answered 'No' to **all** the questions above, go to **Section B**. If you answered 'Yes' to **any** question above, continue to question 5 below.		
5. Is your project covered by a CUREC Approved Procedure (formerly known as "CUREC Protocols")?	Yes ☐	No ☒
If yes, give the specific Approved Procedure number(s):		
If you answered '**Yes'** to **ANY** of questions **1-4**, **and** answered '**No'** to question **5**, **stop completing this checklist and do not submit it for ethical review**. Instead, complete the CUREC 2 application form from the CUREC website, then submit that for ethical review. If you answered '**Yes'** to ANY of questions 1-3, **and** answered '**Yes'** to question 5, go on to **Section B**.		

SECTION B: Contact details and project description	
Contact details:	
1. Principal Investigator **or** supervisor (if student research) (give title and full name)	Professor Jo-Anne Baird
2. Name of student (if student research)	[redacted]
3. Degree programme (if student research), e.g. BA, BSc, MSc, MPhil, DPhil	MSc Educational Assessment
4. Department or Institute name	Education
5. Address for correspondence (if different from above)	15 Norham Gardens, Oxford, OX2 6PY
6. University (not private) e-mail address and telephone number	jo-anne.baird@education.ox.ac.uk
7. Name and status of others taking part in the project (e.g. third year undergraduate; postdoctoral research assistant)	n/a

Project description:	
8. Title of research project	What are our key assessment principles after a summer without assessment?
9. List of location(s) where project will be conducted	Cardiff, Wales, UK Oxford, Wales, UK
10. If your research involves overseas fieldwork or travel and your department requires a travel risk assessment, will you have completed and returned a risk assessment form beforehand? (This must be approved by your department before you travel. If you are travelling overseas, you are strongly advised to take out University travel insurance.) Please also address any physical or psychological risks for Oxford researchers and local fieldworkers in question 16 below and discuss with your Safety Officer.	Yes ☐ No ☐ Not required in this instance ☒
11. Anticipated duration of overall research project	12 months
12. Anticipated start and end dates of the part of the research project involving human participants and/or personal data	From: (07/03/2021) To: (16/09/2021) Note: You will need ethics approval before you start your research. CUREC 1As may take up to **30 days** to process. Retrospective ethics approval **cannot** be granted.

In the case of international or collaborative research, will you submit or have you submitted this project for ethical review or consideration elsewhere (e.g. collaborator's/ local ethics committee, or other local approval)?	Yes ☒ No ☐ If 'Yes', please attach ethics or other approvals and give more details below. If 'No', please explain your reasons below. Please also refer to the Best Practice Guidance on Ethical Review of social-sciences based research conducted outside the UK (BPG 16), which includes an Ethics Issues Checklist for International Research (Appendix A)
Please supply further details in response to question 12b here. WJEC – the suppliers to the quantitative data used for the research – provided approval for this project (subject to CUREC approval) on 23rd February 2021.	
13. External organisation funding the research (if applicable)	WJEC
14. Brief description of research (about 150 words) in lay language. a) When describing the research, include your methodology, how you are applying professional guidelines, and the use to which results/ data will be put. Please also declare any conflicts of interest here.	
The aim of the study is to explore the short- and (potential) long-term effects of the policies in relation to summer 2020 general qualification examination series in Wales and England, with respect to the key assessment principles of validity, reliability, fairness and comparability. More specifically, I intend to evaluate how the various approaches taken to grading prioritised and balanced these principles, using an analysis of documentary evidence. In support of this, given the focus on centre assessment grades and alleged 'downgrading' last summer, I wish to analyse grade data to evaluate the grades submitted by schools and colleges. This will enable a discussion of these grades in each of the approaches set out in the previous section. The study therefore takes a mixed-methods approach. The intention is that the research will inform contemporary debates on the key principles of assessment, how they are and should be balanced, and how they inform the design of future high-stakes assessment in the UK and beyond. The research will be conducted in line with British Educational Research Association guidelines.	
b) Description of participants and how you will obtain informed consent to take part in the research.	
i. Description of participants and your criteria for inclusion/ exclusion.	
Students who received an A level result from WJEC in 2018, 2019 and 2020. Data from 2018 and 2019 will be used as a basis for analysing grades from 2020. A level results in 2018, 2019 and 2020. At the point of analysis, it may be necessary to reduce the focus slightly (to the predominant age group, or Wales only, or those with banked grades at AS level) in order to make the analysis clearly in light of its specific purpose.	
ii. Your method(s) of recruitment.	
No recruitment required, as the proposed analysis is on secondary data (see Section 18i-l).	

iii. Your processes for obtaining informed consent from participants.

Not applicable (see Section 15).	
Please attach separate supporting documents (in Word), if appropriate, for your research (English language versions only). Tick those you are submitting below. If appropriate supporting documents are not submitted, you will be asked to provide these separately, which may delay the ethical review process.	
Recruitment and advertisement material (e.g. a poster, social media recruitment text, or brief invitation letter/ email).	☐
Information for participants to read (or hear) before they agree to take part (e.g. written information or, if applicable, an outline oral information script).	☐
A document to record informed consent. Templates for written consent forms and/ or oral information scripts (in case of an oral consent process) are available from the CUREC website.	☐
Questions to be asked of participants (e.g. interview questions, or a preliminary scope of questions, or a sample questionnaire).	☐
(If applicable) debriefing document after participants have taken part.	☐
Please add any further details. If you feel the above approaches are not appropriate for your study, provide details on how you will obtain consent from participants Please answer question 15 if you cannot obtain informed consent.	
Informed consent cannot be obtained (see subsequent sections).	
15. If you cannot obtain informed consent from participants according to CUREC guidelines and good practice in your discipline, please give a brief explanation and justification of this decision below.	
The secondary data set is very large (anticipated to be over 100,000 records in total) and obtaining informed consent from all participants is not feasible. Data will be anonymised by a WJEC member of staff prior to sharing with the student named above, with only a limited number of personal characteristics (age at 31st August in the year of examination; gender). The documentary material is all publicly available.	
16. What are the ethical issues connected with your research and what steps have you taken to address them? Please do not answer 'none'. We need to see that you have identified potential ethical issues with respect to your research and have taken steps to address them. If applicable, please address:	
a) Participant burdens and/or risks	
As this is secondary analysis of anonymised data, there is no direct burden and/or subsequent risk to research subjects. There is one indirect risk, as the analysis may determine that grades were implausibly high in light of the analysis, at least to some extent. This is already the subject of significant public debate and controversy, and the intention behind the research is to add nuance and new perspectives to that debate, rather than to undermine the awarded grades. As such the main mitigation is to ensure that the analysis is presented fairly and in the difficult context in which the grades were provided.	
b) Your own physical and psychological safety as a researcher or of fieldworkers you may employ (see the University's and Social Science Division's Safety in Fieldwork guidance)	

No fieldwork will be undertaken for this research, so there are no associated physical safety risks. The student is a full-time employee of the organisation whose data is the subject of quantitative analysis, and given the public controversies around the approaches taken to grading in 2020, the documentary evidence being collected is also likely to contain statements which are critical of the work of the student in a professional capacity.

There is therefore a potential emotional impact resulting from the research, which will be managed via open and reflective discussions with the supervisor, who has direct experience of managing similar experiences in a professional environment.

c) Data protection/ confidentiality (also see Section 18).

Although the quantitative data will be anonymised, there is a small risk that individuals could be identified on the basis of the limited personal characteristics data included in the data set, together with the grades and marks awarded to them. Further details on the mitigations for this risk are set out in Section 18.

Any other ethical issues. For more guidance on ethical issues, please see http://researchsupport.admin.ox.ac.uk/governance/ethics/resources

No other ethical issues to note.

17. Will your research involve discussing sensitive issues? This could be information relating to race or ethnic origin, political opinions, religious beliefs, physical/ mental health, trade union membership, sexual life or criminal activities. If you answered 'Yes', make sure you include some supporting information (as directed in **section 14b** above, showing the range of questions covering these issues.	Yes ☐	No ☒

Please provide further details:
n/a

18. Management and handling of personal and other research data

All information provided by participants is considered research data for the purpose of this form. Any research data from which participants can be identified is known as personal data; any personal data which is sensitive is considered special category data.

Management of personal data, either directly or via a third party, must comply with the requirements of the General Data Protection Regulation (GDPR) and the Data Protection Act 2018, as set out in the University's Guidance on Data Protection and Research. In answering the questions below, please also consider the points raised in the Data Protection Checklist and whether, for higher-risk data processing, a separate Data Protection Impact Assessment (DPIA) may also be required for the research. Advice on research data management and security is available from Research Data Oxford and your local IT department. Advice on data protection is available from information.compliance@admin.ox.ac.uk.

a) Please mark 'X' against the data you will collect for your research

Consent records (written consent forms, audio-recorded consent, assent forms (for research involving minors) including participant name)	☐
Online consent (may be anonymous)	☐
Opt-out forms	☐
Contact details for research purposes only (destroyed when no longer needed for this research)	☐
Contact details kept for future studies	☐
Audio recordings (preferably using PIN-protected audio recorder and stored on device's hard drive)	☐
Video recordings	☐
Transcript of audio/ video recordings	☐
Photographs	☐
Task results (e.g. paper/ online tasks, diary completion)	☐
Questionnaire answers	☐
Field notes	☐
Other (please specify below)	☒

- Public statements from key stakeholders engaged in debates regarding general qualification awarding in 2020 and 2021.
- AS and A-level grades and marks awarded in summer 2018, summer 2019 and summer 2020 examination series (Wales-regulated WJEC qualifications only)
- The examination centre number, gender and age of those in receipt of an examination result in those years.

b) For each of the types of data selected above, state how this will be physically transferred from where it is collected to a local secure storage site (and backed up as necessary). This includes paper records and data captured electronically.

The data that will be used to compile the final data file for analysis is already held securely on WJEC's electronic storage. The data file for analysis will be transferred to a separate location within WJEC filespace (see section c) and access limited to the researcher for the purpose of completing the research. No data will therefore leave WJEC storage.

c) How and where will each type of data be stored during the research (until the end of all participant involvement)? Describe the arrangements for ensuring confidentiality, i.e. location of storage (e.g. Nexus 365 OneDrive for Business, SharePoint), security arrangements and de-identification of such data. Do not store unencrypted data in freely available cloud services or unprotected USB drives.		
The data Richard will use for this research will be stored securely within a specific folder on WJEC's OneDrive and will be accessed only from an approved device which satisfies encryption at rest requirements, and has multi-factor authentication and protective monitoring in place.		
d) Will you use a unique participant number on research data instead of a participant name? If yes, state whether or not you will retain a list of participant names against numbers (i.e. pseudonymisation via a linkage list). Where will the list be stored, and when will it be destroyed?		
The requested quantitative data files will contain a pseudo-anonymised key at individual and examination centre (school/college) level. This will ensure that the analysis can link individuals together (for example, if they have taken multiple qualifications), and allow for a consideration of patterns at the centre level. No identifying list will be provided or stored by WJEC, by agreement.		
e) Who will have access to the research data?		
Once created, the data will only be accessible to the student. A copy of the data file and the script used to create the file will be kept securely by WJEC for three years, starting from the date on which the data is shared with the student.		
f) If research data is to be shared with another organisation, how will it be transferred/ disclosed securely?		
n/a		
g) When and how will identifiable data (including audio/ video recordings & photos) be destroyed or deleted? Note: Records of consent should be retained for a minimum of three years after publication or public release. Some funders may require longer periods. If you wish to retain contact details in order to re-approach participants about future studies, you must detail this in information provided to them and obtain specific consent for this.		
The quantitative data file will be kept for a period of three years by WJEC, starting from the date on which the data is shared with the student. After this date, the data file will be destroyed (although it is anticipated that, in line with fair processing, the data files used to produce the file of the student will remain on file for future use).		
h) Please confirm that you will store other research data safely for at least 3 years after final publication or public release and adhere to any additional research funder policies. For more information about the University policies, please see the University's web pages on research data management.	Yes ☒	No ☐
If 'Yes', please give details of who will store the data and on storage format, location and security. Note that open science is encouraged. If 'No', please provide further details below.		

Discussions with interested parties for the sharing of the requested data and related files for research purposes are ongoing, in line with similar conversations in England (see https://ofqual.blog.gov.uk/2020/12/21/ofqual-and-data-sharing/). The proposal presented here is without prejudice in relation to those discussions. Any release of the files used for the purposes of this research would be considered in that wider context.

i) Does your research involve the use of secondary (i.e. previously collected) data? Common sources of secondary data include censuses, information collected by government departments, organisational records and data that was originally collected for other research purposes. (If "No", please go to question 19.)	Yes ☒	No ☐
j) Do you have data access agreements for the use of this secondary data? (If so, please attach these.) The following document is attached: • Data access agreement with WJEC for the purpose of undertaking the research.	Yes ☒	No ☐
k) Is your use of this secondary data compatible with what the data subjects/ participants agreed that their data should be used for?	Yes ☒	No ☐
l) Could this data be linked back to an individual or individuals? If yes, address how securely any personally identifiable data will be transferred to you, and where and for how long it will be stored during or after the research. Who will have access to it?	Yes ☒	No ☐

Although the quantitative data will be anonymised, there is a small risk that individuals could be identified on the basis of the limited personal characteristics data included in the data set, together with the grades and marks awarded to them. The risks of individual exposure are significantly reduced by the security placed around the data:

- The data file for analysis will be transferred to a separate location within WJEC filespace and access limited to the researcher for the purpose of completing the research. No data will therefore leave WJEC storage.
- The requested quantitative data files will be pseudo-anonymised to ensure that the analysis can be completed while making it much more difficult to link the data to any other data held by WJEC.
- A copy of the data file and the script used to create the file will be kept securely by WJEC for three years, starting from the date on which the data is shared with the student. After this date, the data file will be destroyed (although it is anticipated that, in line with fair processing, the data files used to produce the file of the student will remain on file for future use).

19. Publication and dissemination of research data and outputs.
How will you disseminate and feedback project outcomes at the end of the research?

The results of the analysis of this data will be included in the student researcher's MSc dissertation, as well as any publications resulting from the study.

SECTION C: Methods and procedures to be used	
1. Please indicate the methods to be used:	**Mark 'X'**
a) Analysis of existing records	☒
b) Snowball sampling (recruiting through contacts of existing participants)	☐
c) Use of casual or local workers e.g. interpreters	☐
d) Participant observation	☐
e) Covert observation	☐
f) Observation of specific organisational practices	☐
g) Participant completes questionnaire in hard copy	☐
h) Participant completes online questionnaire or other online task	☐
i) Using social media	☐
j) Participant performs paper and pencil task	☐
k) Participant performs verbal or aural task (e.g. for linguistic study)	☐
l) Focus group	☐
m) Interview	☐
n) Audio recording of participant (you will generally need specific consent from participants for this)	☐
o) Video recording of participant (you will generally need specific consent from participants for this)	☐
p) Photography of participant (you will generally need specific consent from participants for this)	☐
q) Others (please specify below)	☐
Please ensure you have addressed any potential ethical issues related to these methods in question B14 and in your Participant Information Sheet.	

SECTION D: Professional guidelines and training		
In this section, please mark 'X' against at least one of the following professional guidelines you aim to adhere to. You should use the principles listed in your chosen guideline(s) in conducting your own research. This is not an exhaustive list.		
Research specialism/ methodology	**Association and guidance document**	**Please mark 'X'**
Anthropology	Association of Social Anthropologists of the UK and Commonwealth	☐
Computer Sciences	ACM Code of Ethics and Professional Conduct	☐
Criminology	British Society of Criminology	☐
Education	British Educational Research Association Ethical Guidelines for Educational Research	☒
Geography	Association of American Geographers Statement on Professional Ethics	☐
History	Oral History Society of the UK Ethical Guidelines	☐
Internet-based Research	British Psychological Society: Conducting Research on the Internet Association of Internet Researchers Ethics Guide ACM Code of Ethics and Professional Conduct Association of Internet Researchers (AoIR) Also see CUREC's Best Practice Guidance on internet-based research	☐
Law (Socio-Legal)	Socio-Legal Studies Association: Statement of Principles of Ethical Research	☐
Management	Academy of Management's Professional Code of Ethics	☐
Political Science	American Political Science Association (APSA) Guide to Professional Ethics in Political Science	☐
Politics	Political Studies Association. Guidelines for Good Professional Conduct	☐
Psychology	British Psychological Society Code of Ethics and Conduct	☐
Social Research	Social Research Association: Ethical Guidelines	☐
Sociology	The British Sociological Association: Statement of Ethical Practice	☐
Visual Research	ESRC National Centre for Research Methods Review Paper: Visual Ethics: Ethical Issues in Visual Research	☐
Other professional guidelines. Please specify the other guidelines used here:		☐

Please indicate what training in research ethics (or research methodology) the researchers involved with this study have received, e.g. the title of the course and date completed (online training available at http://researchsupport.admin.ox.ac.uk/support/training/ethics), or discussions between researchers and supervisors, if applicable.

- Postgraduate Diploma in Social Science Research Methods (Cardiff University, 2005-2007)
- MSc Educational Assessment dissertation training week (Department of Education, September 2020)

SECTION E: Signatures or email endorsements

The SSH IDREC Secretariat accepts either option below. If you have a DREC, check which option it prefers.

Option 1: Email confirmation from a University of Oxford email address can be accepted. Separate emails should come from each of the relevant signatories as outlined below, indicating acceptance of the relevant responsibilities. Pasted images of signatures cannot be accepted.

Option 2: Handwritten signatures. Please scan them and the rest of the checklist pages to create a single PDF document and email to us.

Please ensure this checklist is signed by:

For staff research:
1. Principal Investigator
2. Head of Department (or nominee)

For student research:
1. Principal Investigator (project supervisor)
2. Head of Department (or nominee)
3. Student researcher

Principal Investigator signature/ supervisor signature (if student research)

I understand my responsibilities as Principal Investigator as outlined in the CUREC glossary and guidance on the CUREC website. I declare that the answers above accurately describe the research as presently designed, and that a new checklist will be submitted should the research design change in a way which would alter any of the above responses so as to require completion of CUREC 2 (involving full scrutiny by an IDREC). I will inform the relevant IDREC if I cease to be the principal investigator on this project and supply the name and contact details of my successor if appropriate.

Signature (or email endorsement using the above declaration): _____

Print name: _____ **Date**: _____

Departmental endorsement signature

I have read the research project application named above. On the basis of the information available to me, I:

- consider the PI and student researcher (if applicable) to be aware of their ethical responsibilities in regard to the ethical issues associated with this research;
- am satisfied that: the proposed project design and scientific methodology are sound; the project has been subject to appropriate peer review; and is likely to contribute to existing knowledge and/or to the education and training of the researcher(s) and that it is in the public interest.

Signed by Head of Department or nominee (example nominees for student research include the Director of Graduate Studies/ Director of Undergraduate Studies):

Signature (or email endorsement using the above declaration): _____

Print name: _____ **Date**: _____

Student signature (if student research)

I understand the questions and answers that have been entered above describing the research, and I will ensure that my practice in this research complies with these answers, subject to any modifications made by the principal investigator properly authorised by the CUREC system.

Signature by student (or email endorsement using the above declaration): (email)

Print name: ___[redacted]_____ **Date**: 23rd February 2021_____

Qualifications & Assessment

WJEC, 245 Western Avenue, Cardiff CF5 2YX
029 2026 5000 exams@wjec.co.uk www.wjec.co.uk

Part of WJEC

☎ Direct / Uniongyrchol
029 2026 5350
elaine.carlile@wjec.co.uk

Our reference / Ein cyfeirnod

Elaine Carlile

Date / Dyddiad

16/09/2021

Approval of research project – [redacted] - MSc in Educational Assessment

WJEC has fully supported XXXXXXX in completing the MSc course in educational assessment at Oxford University. The work XXXXXXX has undertaken during the course of his studies has benefitted WJEC in many ways, and we believe the research he will undertake whilst completing his disseration entitled *What are our key assessment principles after a summer without assessment?* will aid WJEC's understanding of the key issues arising from the alternative arrangements put in place following the cancelling of exams in the summer 2020 series. We consider the research will also provide insights into how assessments might be developed for qualifictions to meet the needs of the new curriculum in Wales.

WJEC agrees to share data with [redacted], as agreed in the attached data agreement.

The data [redacted] will use for this research will be stored securely on WJEC's OneDrive and will be accessed only from an approved device which will satisfy encryption at rest requirements, and has multi-factor authentication and protective monitoring in place.

Elaine Carlile
Director of Qualifications and Assessment

WJEC Data Sharing Agreement for Research

WJEC agrees to share data with [redacted], as outlined below, to aid the research he is undertaking in completing his dissertation entitled *What are our key assessment principles after a summer without assessment?* as part of the MSc he is completing in educational assessment.

In order to carry out his research, [redacted] will use WJEC data from 2018 and 2019 series as a basis for analysing grades from 2020. The data will be anonymised by a WJEC member of staff prior to sharing with [redacted], with only a limited number of personal characteristics (age at 31st August in the year of examination; gender). The documentary material is all publicly available.

The data that will be used to compile the final data file for analysis is already held securely on WJEC's electronic storage. The data file for analysis will be transferred to a separate location within WJEC filespace and access limited to the researcher for the purpose of completing the research. No data will leave WJEC storage nor will it be shared outside WJEC.

The requested quantitative data files will contain a pseudo-anonymised key at individual and examination centre (school/college) level. This will ensure that the analysis can link individuals together (for example, if they have taken multiple qualifications), and allow for a consideration of patterns at the centre level. No identifying list will be provided or stored by WJEC.

Once created, the data will only be accessible to [redacted]. A copy of the data file and the script used to create the file will be kept securely by WJEC for three years, starting from the date on which the data is shared. After this date, the data file will be destroyed (although it is anticipated that, in line with fair processing, the data files used to produce the file of the student will remain on file for future use.

By signing this document, I, [redacted], agree that:

a) the data provided by WJEC will be used only for research purposes as outlined above
b) the data provided will not be shared with any other organisation or body
c) all data will be provided by WJEC in good faith as being accurate
d) due acknowledgement must be made to WJEC in the final dissertation
e) the dissertation must include a clear statement to indicate the ideas and opinions expressed in the document are those of the writer
f) the anonymity of students and institutions must be preserved
g) a copy of the draft dissertation must be submitted to WJEC for scrutiny prior to the final version being completed, and a copy of the final version made available to WJEC
h) should WJEC have serious reservations about the use and/or interpretation of its data in the dissertation, whether these reservations are concerned with the methodology used, the findings or any other matter, the dissertation should include a preface indicating the WJEC's reservations and reasons for this
i) the use of any data must be consistent with the principles of the Data Protection Act.

Signed: [redacted] Date: 23 February 2021

Signed:  Date: 23 February 2021

Elaine Carlile, Director of Qualifications and Assessment, on behalf of WJEC