Cambridge Assessment

Education Assessment Research Seminar

Online dissemination of work by leading UK education assessment researchers

The online seminar took place on 12 and 13 October 2020 and was an opportunity, with the permission of the presenters, to widen the audience to other education researchers from university faculties, international education assessment researchers and the like. Please bear in mind that the research was created for the UK context.

Please find below a synopsis of each presentation and answers from the presenters to questions raised in the online Q&A.

Session 1 - Vocational Qualifications

(Chair: Jackie Greatorex)

Watch the session.

Vocational qualifications at Key Stage 4 and Key Stage 5: who takes them and how they fit into students' programmes of study

Sylvia Vitello and Carmen Vidal Rodeiro, Cambridge Assessment

Following major reform to vocational education for 14–19 year olds in England, this research aimed to obtain an up-to-date and comprehensive view of vocational uptake with regard to candidates' characteristics and programmes of study.

Descriptive statistics and regression analyses were conducted to understand;

(1) the balance and types of Vocational Qualifications within students' programmes of study; and

(2) the demographic and educational characteristics (e.g., gender, prior attainment, income-related deprivation) of students with Vocational Qualifications.

Analyses were conducted for Department for Education's categories of qualifications (e.g., Technical Awards, Applied Generals, Technical Levels) and educational pathways (e.g., academic only; mixed; vocational only).

The report can be found on the Cambridge Assessment website.

Were you surprised to see such a clear split between vocational and academic pathways at level 3?

Yes, at first we also found it surprising to see that the vast majority of candidates at level 3 were either on an academic-only or a vocational-only pathway (when considering only their level 3 qualifications). We were expecting to find a larger proportion of students on some kind of mixed pathway, given the recent increases in candidates entering university with at least one vocational qualification. However, we have to remember that students take level 3 qualifications not only for university entry but for all sort of other reasons, such as going directly into employment. So, perhaps, it's not as surprising as first thought that a large proportion of students take vocational qualifications only, without any academic qualifications at all, especially if they are not intending to enter university. The fact that a large percentage of students were on a vocational-only pathway highlighted to us the importance of having vocational qualifications at level 3 that can comprise a students' full programme of study.

Do you think it is significant that a quarter of learners starting Key Stage 5 are not on a level 3 programme?

Definitely. This finding emphasises how important it is that Key Stage 5 provision and reforms to qualifications for this key stage do not focus only on level 3 academic or level 3 vocational qualifications. Key Stage 5 education needs to provide useful level 2 qualifications (academic and vocational) or other learning programmes which can support students who are not yet ready for level 3. This questions the extent to which it is appropriate to have only one Department for

Education-approved category of level 2 vocational qualifications (Technical Certificates) for learners at Key Stage 5.

One of the Department for Education requirements for vocational qualifications to be in performance tables is that they need to support progression–have you looked at the progression from the different pathways at Key Stage 4/5 to further study?

Yes, we have done some research looking at progression from the different Key Stage 4 pathways/qualifications to Key Stage 5 and at progression from one particular type of Key Stage 5 Vocational Qualification (Level 3 Cambridge Technicals) to Higher Education.

Regarding progression from Key Stage 4, the results of our research show that, overall, mostly vocational and vocational-only pathways do not close students' options at Key Stage 5, as candidates were either able to continue in a vocational programme or to progress to different pathways, including academic-only and mostly academic pathways. We are in the process of writing up a journal article with the main findings relating to progression from vocational qualifications at Key Stage 4.

A link will be on the Cambridge Assessment website when it becomes available.

And, regarding progression from Key Stage 5, our research revealed a positive message for level 3 Cambridge Technicals: there is good progression to university and, once candidates have applied, the offer and acceptance rates are high. This shows that Cambridge Technicals are a valuable means of accessing Higher Education and fulfil the requirements for Technical Levels and Applied General qualifications to be included in the 16 to 19 performance tables. A report covering our research on progression from Cambridge Technicals to Higher Education can be found here.

As you have shown us, many students seem to take a mixture of vocational and academic qualifications. Did you notice if vocational and academic qualifications support different subject needs or if there is an overlap in the subjects?

Overall, our research suggested that the Department for Education-approved Vocational Qualifications were covering some subject gaps that academic qualifications (such as GCSEs or AS and A Levels) did not cover. This could be an indication that vocational qualifications are being used to fill subject gaps in the curriculum.

However, there was some overlap in certain subject areas. This overlap could be interpreted in different ways: for example, it could indicate that the different qualifications cover different aspects of a subject area, or that candidates have obtained two qualifications just by learning one set of content. As an example, we found some degree of candidate overlap for Art and Design at Key Stage 4, but Art and Design is quite a broad area of study (i.e., it includes, for example, Graphics, Photography, Textiles, Fine Art, 3D Studies, etc.) so candidates could be taking completely different qualifications. You can find results on subject coverage in our online report.

The data is three years old: is the pattern still the same with some of these Vocational Qualifications now more established in centres and older version removed?

At the time we started the research, the 2017 data was the most recent data available from Department for Education and the reforms to Vocational Qualifications were starting to have an effect. In fact, when the carried out the research there was a mixed economy of legacy and Department for Education-approved Vocational Qualifications being taught at schools.

The reforms to Vocational Qualifications have changed the landscape of vocational education at *secondary level education in these last few years and this has not occurred in isolation but within the context of wider reforms to academic qualifications and to the secondary education system more generally. As a result, in the last few years provision and uptake of these qualifications might certainly have changed.*

To see if there has been changes in the patterns we found in our research, we are planning to carry out some further research and we are in the process of writing a data request to Department for Education to get data on full cohorts of students at the end of Key Stage 4 in 2018 and 2019 (linked to their Key Stage 5 results two years later).

Evaluating qualifications: applying a validation framework to BTEC Level 3 Nationals

Hayley Dalton and Kevin Mason, Pearson

Last year, we presented a work-in-progress development of a framework for evaluating GCSE Mathematics in order to communicate validity of Pearson UK qualifications in a way that is robust and accessible. This built on a similar process looking at validity by Pearson colleagues in the US, with reference to the work following Shaw and Crisp (2012) and Newton (2017).

For our work on BTEC Level 3, we have started at the final part of that framework and focused on the purpose of the qualification, and how the qualification has been designed to meet that purpose and early indications of outcomes. We concentrate on BTEC Level 3 Nationals in Health and Social Care in order to understand and demonstrate the ways in which this, and similar qualifications in the BTEC Nationals suite, support learners through their post-16 education. We investigate how the ways in which the qualification is designed and delivered prepares learners for progression into sector-related higher education, especially in a sector with skills shortages; also how learners themselves reflect on their own future hopes for employment.

How much confidence do you have that self-reported employment hopes will match up with real employment prospects?

The Employment Hope Scale is just one part of the jigsaw of data we are using as part of the framework. It is not intended to be a proxy for actual employment progression data, so we wouldn’t necessarily expect it to match up closely. It is helpful in indicating which BTEC sectors/subjects are preparing learners for entry to the labour market effectively (or are perceived to be by learners). We also intend to contrast it with other leading indicators to ascertain how applied general qualifications support choice and career pathways compared to Careers Education, Information, Advice and Guidance or any other types of intervention.

How did you recruit the students to the survey?

Through our network of Quality Nominees in every centre. Incentives were offered to participate, and we employed extensive phone chasing throughout.

How do you plan to collate employment impact data in the long term for any given sector?

We are going to employ a case study model on sectors where we feel employment data would be most valuable. Current thinking is that we will do some deep dives into selected occupations to map the content of the qualifications alongside the skills profiles for occupational areas with known vacancy gaps. This will focus of outcomes rather than impacts in the first instance.

Data is 3 years old: is the pattern still the same with some of these Vocational Qualifications now more established in centres and older version removed?

Many patterns have stayed the same with respect to learner experience in particular - possibly because much of this is linked to centre activity as opposed to the content or assessment of the qualification. Where we have seen changes with the introduction of the new qualifications it is more linked to access and how centres are recruiting and supporting learners on the programme.

Exploring internal assessment in national Vocational and Technical Qualifications

Charlotte Lockyer and Stuart Cadwallader, Ofqual

Internal assessment is widely used in Vocational and Technical Qualifications (VTQs) to assess practical skills and competencies which can be more difficult to assess through an external exam. Although it is widespread, there is little published research on how internal assessment works on the ground in VTQs.

During 2018 and 2019, Ofqual carried out a qualitative study to establish a contemporary overview of practice in the internal assessment of six case study VTQs. Two qualifications were sampled from each of the three sector subject areas (Construction, Hair & Beauty and IT). We aimed to use these qualifications to better understand differing practices around internal assessment in a range of contexts. We also aimed to explore whether there are systemic vulnerabilities which might threaten the validity or reliability of assessment in VTQs.

Our study involved a series of in-depth interviews with teacher-assessors delivering the six qualifications. We carried out interviews with 45 assessors in 21 centres. The scope of our interviews was wide, covering summative task setting; task taking; learner support; how assessors make assessment judgements; and quality assurance. We deliberately adopted this holistic approach in order to develop a contextualised understanding of how the system works as a whole.

Our study served to highlight just how diverse, complex and dynamic the VTQ sector is. The six VTQs in our sample were delivered to a different learner base in different settings and had very different purposes. Despite this diversity, we were still able to identify some common themes in the qualitative data. Our presentation will discuss some of these themes in more detail.

Did it seem as though assessors' practice had frequently had to change (for the softly vocational qualifications) or that the frequent change had almost de-sensitised them to change (and they tended not to react too much to another round of changes)?

I think there was a bit of both. Logistically there is a lot assessors who have to change when centres deliver new courses, and there is a huge amount of planning that goes into this. In this way, frequent change could be a source of real frustration for some assessors. But in other ways, thinking specifically about how assessors internalise a new set of assessment criteria for example, there did seem to be an element of desensitisation to change. Assessors who had delivered a series of quite similar courses in the past often appeared to have developed an internal conception of the assessment standard. So whilst they would absolutely be guided by the new assessment criteria, they would view this through the lens of this past experience of qualification delivery, and this more 'constant', stable internal conception of what an appropriate standard in a certain criteria might be. Whilst not a 'softly vocational qualification', one hairdressing assessor summed this up when she said "What I always keep in my mind is that 'shampooing' is 'shampooing' whatever you call it. Blow drying is blow drying. Setting is setting. So it doesn't really matter too much about individual [qualification] numbers."

Did you speak to awarding organisations to discuss why certain approaches were being taken?

No. In this piece of work we decided to focus on assessors' experiences and perceptions of the system. So whilst we shared our findings with awarding organisations to ensure there were no inaccuracies in our reporting, we didn't then go on to explore their rationale for adopting certain *approaches. As an exploratory piece of research, this was intended to be a starting point for various pieces of follow up work. We have been able to carry out some of this follow up work since, working with awarding organisations to understand the aims and approaches to external quality assurance, for example. However, some of this has had to be placed on hold as a result of the COVID-19 pandemic.*

Drivers of choice for Vocational and Technical Qualifications

Stephen Price, Ofqual

The reasons that centres choose to offer particular Vocational and Technical Qualifications (VTQs) to 16—19 year olds has received relatively little attention despite these qualifications forming a large share of the qualifications market. Ofqual surveyed centres in England on the drivers that contribute towards the choices of VTQs they offer as well as the awarding organisations they buy from. The survey comprised three major sections:

1) background information;

2) drivers of choice of VTQ offerings and factors affecting AO of choice; and

3) how centres manage their offerings over time.

The survey received 503 responses representing centres categorised into six main types and the data were weighted to yield a representative sample.

Centres take a broad view when it comes to making choices about VTQ offerings, seeming to balance a range of factors, including those that are principally focused on learners (e.g., their interests and prospects) as well as those focused on the centres themselves (e.g., capacity). Centres seem particularly committed to meeting perceived learners' needs. Schools and colleges differed from each other in their general level of consensus about drivers of choice: colleges appear to be clear that VTQ choices can maximise their student recruitment ability and that their remit is to prepare learners for next steps. On the other hand, there is perhaps less of a clearly defined role for VTQs in schools meaning that priorities, and therefore drivers, will vary more from centre to centre in line with the specific context of a school.

Did centres comment on T-levels and their reasons for engaging or not?

We did not specifically ask about T-levels.

You said you asked respondents to consider existing and new/proposed qualifications. Did you see any difference in the key factors between new and existing?

The differences between the top two existing qualifications and any new or proposed qualifications were remarkably small but it will be interesting to see how things change following the disruption caused by Covid-19.

What makes vocational assessments fair?

Stuart Shaw and Isabel Nisbet, Cambridge Assessment

This presentation will attempt to address how the various senses of fairness (identified for educational assessments) apply in vocational contexts; and, whether there are new senses which are important for vocational assessments.

What practical steps should awarding organisations and regulators consider to reduce unconscious bias?

We argued in the presentation that related concepts of fairness of use include:

- *construct validity–related to the job or profession in question*
- *freedom of bias–including underlying bias in the assessment method, for example, writing about building a wall as opposed to actually building one.*

Some participants in our data set expressed concerns about the quantity of written assessment required in some Vocational Qualifications (e.g., writing about building a wall) the focus being on burden and irrelevance of written assessment. It is considered unfair by many of the respondents that learners who are good at writing are more likely to excel than those who are merely adequate at writing but strong performers in a skills context. Otherwise excellent apprentices are unable to complete their overall assessment because of not passing the requisite 'Essential Skills' qualifications (in communications and application of number) for the apprenticeship framework. Respondents talked about bias against those who are not 'academically-minded' but who are otherwise occupationally skilled/competent in practical roles.

Practical steps taken by awarding organisations would be to seek to significantly reduce the use of written assessments in vocational assessment. There was a perception that qualification design had been too much concerned with having to display the processes described by "describe, explain, analyse" rather than demonstrating practical and technical skills.

Awarding organisations would do well to increase the clarity of their assessment system: enhancing the clarity of the assessment regime would aid fairness for the individual learner. Features designed to increase clarity might include:

- *ensuring that tests are written well with no ambiguity or bias and using the appropriate level of language for the qualification.*
- *securing an appropriate coverage of the specification across assessments, while making sure that any access restrictions to assessments are strictly limited to the inviolable needs of an industry (e.g., case of a licence to practise).*
- *some indicative content to help to exemplify the expectations of the learner.*
- *setting out what special arrangements and reasonable adjustments to be considered.*
- *sharing practice examples between AOs–this could be valuable so that learners would know that they would be treated in a comparable way.*

In the book Is Assessment Fair? we argue for taking a "situated" view when considering the fairness of an assessment–not just looking at it under the technical microscope away from its context. The "isolated" view looks at an assessment in isolation from its context and sets aside extraneous information about the situations in which it was taken and the uses to which the outcomes were put in practice. In contrast to the 'isolated', the "situated" view, considers assessments/tests in the contexts in which they are taken and the outcomes are used. The situated view also allows for some applications of the "retrospective" (In this sense, an *assessment may be unfair if its outcomes are the consequence of unfair/socially unjust actions in the past, such as racial segregation of school education) and "consequential" (In this sense, an assessment could be accused of unfairness if its outcomes might be used as a basis for unfair or unjust actions in the future) senses of fairness identified in the presentation.*

A few caveats relating to the situated view when used in judging the fairness of vocational assessments or in criticising an assessment organisation for unfairness. Awarding organisations may not know the contexts in which vocational assessments will be taken (other than the fact that the same exam will be taken by students in very different circumstances). Also, it is common for different awarding organisations to be responsible for different parts of the life-cycle and context of an assessment and its use. Additionally, test outcomes can sometimes be used in ways that awarding organisations could not be expected to anticipate.

Did your findings show differences according to whether participants were involved more with strongly vocational or softly vocational qualifications?

I think there are elements of fairness buried inside this deeply-felt bundle of emotions. First, some of the negative comments and attitudes to some VQs are not based on any evidence at all–in Isabel's time at Ofqual she was frequently asked about "Mickey Mouse qualifications", in response to which she always politely (she hopes) asked the questioner whether they had looked at the content of any of these qualifications. Frequently they had not. That is unfairness in the retributive sense–the low esteem is not deserved. More important, perhaps, is Jonathan Wolff's concept of equality of "respect-standing". That is a form of relational fairness and is relevant here. We also referred to retrospective fairness, where the low esteem is linked to students' self-perception as failures resulting from being labelled as such throughout their school career. We have been told about unpublished research by Pearson comparing university attitudes to students who come with BTEC qualifications to attitudes to students who come with A Levels. If they are treated differently because of the labels of their qualifications rather than because of the knowledge and skill they bring to the course, then that is unfair (again in a relational sense).

Do you think that parity of esteem is really about fairness–or is it just sour grapes on the part of those who work in vocational education?

The distinction between "stronger" and "softer" is an interesting, necessary and helpful one (Charlotte's presentation). As we understand Charlotte, the more closely vocational qualifications are related to specific job roles the "stronger" the qualification is; the more generic they are, for example, general health & social care (rather than, say, early years child care) then the "softer" they are. In the strong sense, the question of whether Vocational Qualifications faithfully represent an accurate and meaningful measure of what employers and the relevant industry requires becomes increasingly important. Whilst participants responded in substantive ways to the survey questions - drawing extensively on their vast experience of vocational qualifications - awareness of the distinction was not readily manifest in their responses. Despite exemplification in their comments, respondents directed most of their attention to "softer" conceptualisations of Vocational Qualifications. There were some notable exceptions, however. One example, from the perspective of employees/professionals, indicates that there are systemic barriers to those from lower socio-economic groups. For example, the route to qualify as a solicitor (specific job role) is almost exclusively via graduate entry (there is an apprenticeship route) onto the Legal Practice Course–the Vocational Qualification that gains entry to being a regulated solicitor. It is, in the Law Society's own words, 'an expensive course' offered by a range of HEIs. This is well established and accepted. (Incidentally, barristers must pass the Bar Professional Training Course (BPTC) with mini-pupillages and pupillages either side, again an expensive route to qualification.) This distinction has relevance to the question: Fairness for whom? The "strong" sense providing greater transparency for those who employ *others in vocational occupations.*

Session 2 - Accessibility

(Chair: Gill Elliott)

Watch the session.

Creating better tests–students' views on the accessibility of different question design features

Vicki Crisp and Sylwia Macinska, Cambridge Assessment

OCR has recently developed accessibility principles for Science GCSE exam design. This research investigated students' perceptions of question features relating to these principles. Two versions of a test were constructed using versions of questions with or without the accessibility principles applied. Year 11 Science students in four schools attempted the test and 57 students were interviewed. Students were asked how easy to understand they found the questions and how different features affected this. For most of the accessibility principles explored, student views were in line with expectations about effects on accessibility. In a few cases, findings were neutral or mixed.

The report can be found on the Cambridge Assessment website.

Did any of the interviewed students have motor, physical, or visual impairments?

No they didn't. Within the scope of what was possible in the study we knew we weren't going to be able to adequately represent different groups of young people so we didn't make this a focus in our recruitment. However, we did have a small number of students with special needs such as dyslexia amongst the interviewees.

Do you take into account any specific challenges individual students have, or are you focuses on impact for a "typical" student?

In this research we focused on a general group of students including a few with specific needs. We would have liked to compare groups with different needs but we weren't expecting to be able to interview enough students to be able to have enough students with different needs in order to make meaningful comparisons. In past research we have looked at the effects of particular question features for learners with dyslexia (Crisp, Johnson & Novakovic, 2012) and for those with weaker reading skills (Crisp, 2011; 2015).

Crisp, V., Johnson, M. & Novakovic, N. (2012) The effects of features of examination questions on the performance of students with dyslexia. British Educational Research Journal, 38(5), 813-839.

Crisp, V. (2011) Exploring features that affect the difficulty and functioning of science exam questions for those with reading difficulties. Irish Educational Studies, 30(3), 323-343.

Crisp, V. (2015) Exploring the difficulty of mathematics examination questions for weaker readers, Educational Studies, 41(3), 276-292.

Was any consideration given to accessibility paper versus on-screen test formats?

No, I'm afraid this wasn't part of our research but it would certainly be interesting to investigate this. The presentation in this session by Sylwia and Carla reports on some relevant work reviewing literature on the accessibility of on-screen assessment.

The way the actual question was asked in the Chemistry question was slightly different too. You split up the question into a statement and then the question. Did students find this helpful and more accessible? It is something I have wondered about many a time.

This is, indeed, another interesting change that was made when the accessibility principles were applied. Because there were several other accessibility themes to discuss with students for this exam question, we didn't ask them about this issue in this question. However, we did ask them about a similar issue in a different question and the majority of learners felt that the version with statement and then question was more accessible (for more detail see question 7a in the Research Matters article about this research).

Order of multiple choice questions is interesting. Sometimes learners feel drawn to the middle option - did alphabetical order and numerical order sway their confidence in their answers?

We found little evidence that alphabetical order or numerical order swayed learners' confidence in their answers. One student expressed that they didn't like it if they thought the first option in the list was the correct answer, perhaps suggesting an expectation that the first option wouldn't be correct, but this was the only case of any evidence of a tendency towards the middle. For the question where we looked at alphabetical order, the students often commented that they would read all options anyway so the order made no difference. For the question where we looked at numerical order, some students felt that they would have found the correct answer regardless of order, but there were also comments that numerical order helped them find the answer they wanted and helped them avoid picking the wrong one due to expecting numerical order when it was not.

Embedded tools for accessible on-screen assessment: a review of the evidence on effectiveness

Sylwia Macinska and Carla Pastorino, Cambridge Assessment

On-screen assessment offers new opportunities to enhance test accessibility, allowing to easily incorporate a variety of accessibility features into assessment platforms. On the other hand, these innovations may also raise new questions and issues about the suitability, applicability and validity of on-screen tests designed with accessibility in mind. This review aimed to identify the accessibility features (and selected assistive technologies) that may be embedded in the design of on-screen assessments and evaluate the evidence regarding these features' adequacy and effectiveness in supporting candidates with special educational needs or disabilities to inform the development of accessible on-screen assessments.

Over 40 different on-screen tools and technologies were identified as currently being in use to support the needs of learners with a variety of specific learning needs or disabilities. Summaries on the effectiveness of each of these tools were produced drawing on the literature of on-screen assessment, e-learning and user experience design. One crucial finding of this review, however, was that the evidence regarding the effectiveness of many of the tools is often lacking or inconclusive, emphasising the need for more research into the impact that on-screen accessibility tools have on learners' performance.

We are in the process of writing up a journal article with the main findings of this work. A link will be on the Cambridge Assessment website when it becomes available.

Was there any consideration of WCAG 2.1 or the Euro accessibility standard EN 301 549?

We have initially considered using a variety of accessibility guidelines as sources of evidence. However, as many of the guidelines are not based on academic research and contain little or no references, we decided to focus solely on experimental work and summarise evidence directly from the literature. This approach also helped us to ensure that we presented the evidence in its original form rather than report on several guidelines that may have been based on the same study/anecdotal evidence, which could have affected the strength of evidence in support/against a specific accessibility feature.

The test scheduling is fascinating; are you planning to do further research into this?

At present, there are no internal research projects planned looking into test scheduling effects. However, this topic has become more popular in the past five years or so (particularly since the discovery of circadian rhythms received a Nobel Prize in 2017) and we can expect to see more studies coming out on this in the near future.

The use of assistive technology in assessment: the experience of teachers and students

Diana Tonin and Stuart Cadwallader, Ofqual

Awarding organisations are required to design their assessments such that they do not present any unjustifiable barriers to students, regardless of any disability or impairment that a student may have (Ofqual, 2019). One way in which students may need to access assessments is through the use of Assistive Technologies (ATs). Building on input from the Accessibility Consultation Forum, Ofqual conducted a qualitative study to gain a more nuanced understanding of how centres use ATs to support their learners and how students experience using ATs for assessment. We conducted semi-structured interviews with a small but diverse range of teachers, SENCos and students, analysing the data using a thematic approach. Overall, teachers and students feel that AT is helpful in multiple ways and that it levels the playing field by, in most cases, removing construct irrelevant variance. However, barriers to the effective use of ATs were also identified, particularly in relation to grey areas in the guidance and issues with the format of the examination papers provided. These findings provide insights that broaden our understanding of access arrangements and how they may best optimise assessment validity and promote fairness for all candidates.

What scope might there be for assistive technologies to help with mental well-being? Taking remotely proctored assessment is very stressful for students in isolated conditions–at least in an exam hall, stressful though that is, it is a shared experience.

This is a good question, but one that unfortunately lies outside of the scope of this particular project - remote proctoring was not really covered by our participants, possibly because we undertook the interviews prior to the pandemic.

Our findings do suggest that the use of assistive technology changes the experience of undertaking assessment, though the exact nature and impact of that change is usually quite personal to the student. With regard to exam stress, for some the availability of assistive technology was reassuring and reduced their stress, while for others it increased it (e.g., through concerns about it malfunctioning). Of course, sometimes it was a bit of both!

Although it did not come up in this research, you raise an interesting question about the impact of remote assessment, both for those who use assistive tech and those who do not. Very happy to discuss this further, it is certainly an area of interest for Ofqual.

Did any schools state they were successfully delivering non-General Qualifications with assistive technology and on-screen delivery?

Apologies that we only had a short time to touch on this after our discussion. Yes, the schools we spoke to were using assistive technology to support students who were taking a variety of qualifications. They discussed using different arrangements for different subjects, but did not usually mention variation by qualification. Their emphasis, regardless of qualification, tended to be on ensuring that the assistive technology that was used for an assessment was the same as that which the student used in lessons and at home normally.

Impact of access arrangements on performance

Carmen Vidal Rodeiro and Sylwia Macinska, Cambridge Assessment

Access arrangements are reasonable adjustments agreed before an assessment to allow candidates with specific needs (e.g., special education needs; disabilities; temporary injuries) to access the assessment without changing its demands.

This research investigated whether access arrangements provided a level playing field offered an advantage to candidates. To this end, the performance of candidates who completed high stakes examinations with and without access arrangements was compared. To account for group differences that have the potential to affect candidates' performance, candidates were matched on a number of background characteristics.

The results revealed that candidates with and without access arrangements performed similarly, suggesting that the arrangements are working as intended.

We are in the process of writing up a journal article with the main findings of this work. A link will be on the Cambridge Assessment website when it becomes available.

How will this research be shared with schools to inform their practice and approach?

One of the aims of this research was to gather evidence of how access arrangements were working and this was, in part, motivated by Cambridge Assessment International Education's plans to revise their access arrangements provision, procedures/processes and guidance to better address centres' and candidates' needs. Unfortunately, the research was finished and a report submitted to Cambridge International just before the UK went into lockdown and examinations provided by Cambridge International cancelled. The delivery of the June 2020 results became then a priority and this research was not shared with schools at that time. After the November 2020 session, we will discuss with Cambridge International how best to communicate our findings to schools.

Were there any differences in how the access arrangements affected performance across different groups (e.g., school type, gender, deprivation)?

We did not analyse the data in this way so we don't know if there were any differences in how access arrangements affect performance across different groups of students. However, it might be interesting to carry out some exploratory analyses to investigate this issue so we will keep it in mind for our future research in this area.

Do you have any plans to carry out further research in this area?

Definitely - we have a couple of studies in progress.

The first study is a more qualitative piece of work looking at stakeholders' views on access arrangements. In this project, we will continue to work with Cambridge International to explore their teachers' experiences and perceptions of access arrangements and how these vary by country.

The second study looks at access arrangements in the UK and it has two strands. The first strand is a quantitative strand looking at uptake of access arrangements (over time and broken down by centre type, students' education needs, etc.) and their impact on performance. The second strand is a qualitative one, very similar to the work we are doing for Cambridge

International but, in the case of the UK both teachers' and students' views will be sought. We will be focusing on the understanding of the arrangements and their views on the benefits they provide, including usefulness and perceived effectiveness of use.

Leadership in the context of examining: perspectives of assessment specialists

Emma Walland, Cambridge Assessment

Examiner Team Leaders are vital to the assessment process as they play a crucial role in ensuring that high-quality marking is completed according to deadlines. This qualitative study explored the qualities and behaviours perceived to be important for success within this context from the perspectives of Assistant Examiners, senior examiners and OCR staff.

Seventeen in-depth semi-structured interviews were conducted, and the data was combined with comments made by senior examiners via examiner review systems. The data was collectively analysed using thematic content analysis.

The findings aligned well with existing literature in this area. Overall, it was found that the shift to online standardisation and the need for frequent feedback makes this particular leadership context challenging. Successful Team Leaders were viewed as being able to create a positive and supportive team culture despite the lack of face-to-face interaction and the competing demands on their time. Leadership in this context was perceived to be focused around the shared goal of achieving fairness to candidates. Through this, the importance of the Team Leader showing empathy for both candidates and other examiners was highlighted. The findings facilitate a nuanced understanding of leadership in this context which can be compared and contrasted with leadership in other contexts.

Did any interviewees express opinions on the process of how they are remunerated?

No, they did not in my research. They were asked about their experiences with their Team Leaders, rather than broader factors of the system such as remuneration. However, previous work suggests that they are more motivated by their pride and meaning in the role, rather than the financial factors.

Do you think that any Assistant Examiners could become a Team Leader with the right training?

I think training is important to ensure that team leaders are made aware of some issues that some Assistant Examiners may face and what strategies they found helpful to overcome these. In my opinion, I think there needs to be that desire to become a team leader too.

Were there any particular challenges with employing a high proportion of retirees? For example, was there any grammatical snobbishness that needed to be curbed?

No, I didn't come across any such issues with employing retirees in my research, just the potential challenge of mismatches in working patterns, which can be resolved with clear upfront communication.

Will it be possible to use the research findings (and previous research findings) in team leader training or recruitment?

I think the findings are useful for training, for team leaders to be made aware of some issues that some Assistant Examiners may face and what strategies they found helpful.

Did any interviewees express opinions on the process of how they are remunerated?

No, they did not in my research. They were asked about their experiences with their Team Leaders, rather than broader factors of the system such as remuneration. However, previous

work suggests that they are more motivated by their pride and meaning in the role, rather than the financial factors.

Session 3 - Marking

(Chair: Irenka Suto)

Watch the session.

Review of marking–comparing different types of feedback

Antonia Sudkaemper and Breanne Chryst, OCR

In this study, we compared different feedback types during Review of Marking with regards to the quality of reviews and assessors' experience. Specifically, we compared the current approach monitor feedback and potential alternative automatic seeding feedback. During a six-week trial, 45 assessors reviewed 300 scripts in total, whilst receiving each feedback type on 100 scripts before moving on to another feedback type.

The results show that monitor feedback and seeding feedback fared equally well with regards to quality of reviews. Most assessors rated monitor feedback to be the more satisfying type of feedback with regards to their motivation, their performance improvement, helpfulness, and efficiency. However, most assessors were similarly satisfied with seeding feedback.

An additional cost analysis based on rough estimates of implementation and annual costs within OCR showed that seeding feedback would be more cost and resource effective.

Did the 'no feedback' group know that they were not going to get feedback? Did it make them more focused about their marking decisions knowing that there was not going to be any other input?

Yes, the participants knew when they were not receiving any feedback, so it is possible that they were particularly focused during this time. However, from the post-study survey and conversations with the participants throughout the study, my impression was that they were very focused at all times and took their task very seriously, whether they were receiving feedback or not.

Are there any plans to introduce chatbots to provide feedback?

Not chatbots as such, I believe the technology is not advanced enough at this point to make it a less than frustrating experience. However, the automatic seeding feedback would pop-up immediately after a seeding script is completed, and we are considering having a (human) contact person available for assessors to clarify if feedback has been ambiguous.

Was the seeding feedback automatic (pop-up style)?

The feedback in the experiment was not automatic as such, rather the lead researcher would send it manually in a way that was simulating automatic feedback as RM Assessor3 does not have a seeding feedback function (yet). If we go ahead with implementing seeding feedback, we will develop a function to make it automatic though.

How are you planning to do feedback in future exam sessions?

It is undecided at this point. We are contemplating the pros and cons of the different approaches and will make decisions accordingly. It is unlikely there will be any changes for the immediate exam series that are currently in planning.

The examiners who did not have feedback were probably using their prior experience to mark. There are prior research findings suggesting that the type of feedback may not be the biggest factor in determining how well examiners mark, or to put it another way a particular type of feedback is sometimes effective and sometimes not so effective.

Absolutely, I am sure the assessors were drawing on their previous experience whilst marking. However, the results did show that there were statistically significant effects for feedback type. This means that the way the feedback was provided did impact on the quality of results.

I carry out 'Review of Marking' as a senior examiner - did you look at all at the impact of different approaches on those giving the feedback rather than those receiving it?

We did give a post-study survey to the three monitors who were involved in the study, but of course they could only comment on their experience with monitoring, and not with other feedback types. Moreover, they were only three of them which is not big enough a sample to draw meaningful conclusions. It might be interesting to conduct another study focusing on the experience of the monitors in monitor feedback.

Applying different measurement theories to evaluate marker reliability in vocational assessments

Zeeshan Rahman, City and Guilds

A variety of factors that exist in the assessment process, such as markers, questions and learners, can introduce unreliability or error in results given to learners. The larger the error, the less confidence we have in assessment outcomes. This ultimately compromises the validity of assessments and the reputation of assessment organisations. Research can help evaluate the reliability of an assessment, provide important information on its quality, and indicate how it can be improved. City and Guilds carried out several research studies to evaluate marker reliability, which primarily involved multi-marker studies where groups of markers were asked to mark the same learner scripts for vocational examinations such as functional Maths & English, advanced Mathematics, Electronic Communication, Beauty Therapy and Make-up Artistry. The marking was evaluated using different measurement theories i.e., classical test theory (e.g., mean score differences, score correlation, grade agreement), item response theory (e.g., many-facet Rasch measurement model) and generalisability theory (e.g., phi/phi lambda and analysis of variance). The aim was to compare findings based on different approaches with a view to exploring the benefits and limitations of these theories in investigating marker reliability. This paper aims to provide an overview of findings from this research.

Did someone inspect the items highlighted for issues in marking to see which of the various methods and approaches gave better fit to observed potential causes?

Yes, we met with a team of experts responsible for assessment principles, development, quality and/or research after each multi-marker study to review marker performance as well as potential issues with items and mark schemes. In some cases, issues with items and mark schemes were identified when these were reviewed more closely based on the findings from the multi-marker studies. The aim was to either correct these issues and/or use lessons learnt for developing items or mark schemes in the future.

‘Explainability’ of machine learning algorithms and implications for reviews of marking and appeals

Cesare Aloisi, AQA

This presentation considers the challenges exam providers would encounter if post-results services for schools (e.g., review of marking or appeals) were carried out by an algorithm. We focus in particular on the issue of ‘explainability’: how decisions produced by an algorithm would have to be made understandable by humans. In England, current post-results regulations are underpinned by principles such as accountability, public trust and fairness; hence, if the use of machines was permitted in the marking review process, machine-to-human communication would be inevitable. We look at how key concepts such as ‘interpretability’ and ‘transparency’ are currently understood and how they are implemented in practice. We also consider how effective widely-used methods such as LIME or saliency maps would be at dealing with the complexity of post-results services. We move on to focus on the performance and capabilities of general natural language processing algorithms such as ELMo, BERT and OpenAI’s GPT-2. We consider how they could be adopted for post-results purposes. Using actual examinations and student responses, we illustrate the limitations of these algorithms with respect to explainability. Finally, we note that current shortcomings of machine learning algorithms are often related to mark schemes and their interpretations. We therefore conclude that mark scheme research may regain prominence over the next few years.

Do you think that AI will be able to write and review research reports? And if so when? And if not why not?

I use AI as a shorthand for machine learning for instance and that’s a very important distinction, so they can do very few things at the same time. You need to have several engines to do one full thing like review your reports. What we know at the moment is that you can use these systems for example in the medical sector to review medical reports and that’s because they are highly structured and it’s very predictable how the new one will look like. So research reports are quite structured, so in principle it should be possible to train a system to review them.

What I would argue here is that they are getting better and better at making predictions but the explanation is something different. Deep learning systems are very shallow explainers and that would not change when you ask them to review a research report. Even if you had, let’s call them computers and every computer has a task, even if you have one computer looking at the outline and one computer looking at the arguments, one at the data, and you get a whole system–together, they could score your research report, but individually they wouldn’t be able to tell you why they produced a certain score. So we need really to get explanations right, before in my opinion we can continue with improving predictions because they really stick and are very superficial level whereas with a person you can take them from different angles and you can say ‘ok, but what did you mean by that?’ and ‘what do you mean by this’ and with an engine you have to design it to take those questions and none exist.

Could AI work as a 'double-check' - to make sure key words haven't been missed due to human error, and flag assignments for a review? What applications are there for machine learning as a tool to help rather than a replacement for human markers.

I think so, provided that we design machine learning engines for specific purposes and we are clear about those purposes. Machine learning is great for detecting patterns and inconsistencies, even those invisible to humans. Newer systems are trained on billions of documents by big companies and therefore need just some fine-tuning to be used in different contexts. However, they still need to be given clear instructions that have, at least in principle, a clear output.

Therefore, I can see how we could use a collection of machine learning engines to produce a report on grammar, legibility, keywords, textual cohesion and other features in a student essay to assist markers in their role, or to flag assignments for a review. Nevertheless, what we are still missing in terms of research on AI and educational assessment is a link between (a) what we deem important in a candidate's work, (b) a mark scheme that allows rewarding the important parts, (c) how we communicate that to a machine learning engine, and (d) how we can ensure that the engine can talk back to check that it considered only the important parts when making a decision.

Session 4 - Maintaining Standards, part 1

(Chair: Tom Bramley)

Watch the session.

Introducing the simplified pairs method for standard maintaining

Tom Benton, Cambridge Assessment

This presentation will introduce a new method of using comparative judgement to help understand the relative difficulty of two examinations. This may be particularly useful in the context of maintaining examination standards over time. The new method is easier to understand and apply than existing approaches as it does not require the estimation of script quality measures using a Bradley-Terry model or similar. As well as exploring the possible gains in efficiency, the presentation will also show other benefits of the simplified pairs method such as generation of confidence intervals around results, and an enhanced ability to detect and address differences between judges.

The report has now been published on the Cambridge Assessment website.

How do you determine the high and low scores for the Comparative Judgement exercise that are excluded? What parameters do you use?

Current practice is to restrict to scripts with scores between 20% and 90% of the available total. There is a little flexibility if there's a shortage of scripts (so can use 15%-95% if necessary) but the 20%-90% range seems to work well as a guideline.

Do you have any thoughts on how this could be adapted to low entry syllabuses (where you won't have hundreds of scripts available)?

Just a matter of reusing scripts across different pairs as much as is necessary. You may also use more comparisons per script (and more in total) to compensate for having fewer scripts. The analysis method stays the same (with a slightly adjustment to how confidence intervals are calculated).

Are the marks awarded left on the scripts or removed?

They're removed. Scripts are presented to judges without any visible marks or annotations.

With Comparative Judgement being used to improve and quality assuring item creation, would that practice give more stakeholder confidence regarding year-on-year changes to scripts?

It's hard to say. Stakeholders are becoming more comfortable with Comparative Judgement in lots of contexts and seem to like it. However, it's difficult to know how well it would hold up if were to lead to an unpopular recommendation (e.g., higher grade boundaries and lower pass rates).

How does this work if one test is "easier" at the top end but harder at the bottom end - which is likely to be reality?

Didn't have much to time to talk about it in the presentation but the technique does extend to give a mapping across the entire score range on each test (slide 13). It's just based on logistic regression of judges decisions on both test scores rather than only using the difference between them. This approach was also tested in our real experiments (slides 15 and 16). The technique correctly identified for the Maths tests that most of the differences in difficulty was at the top of the score range, whereas for the English Literature test the difference in difficulty was mostly at the lower end.

How do you get the results if you do not use a Bradley-Terry model?

We directly link scores on each test to the judges' decisions. The easiest way to do this is to look at the relationship with difference in mark (slide 9) but it can be extended using logistic regression of judges' decisions on the marks on each separate test version.

OCR trials of comparative judgement for awarding: reliability, utility and validity considerations

Sarah Hughes and Tony Leech, OCR

Practical and technical issues with grade boundary determination procedures for qualifications in England are substantial, especially in the light of the unique circumstances for standard maintaining in 2020, 2021 and perhaps thereafter. Comparative judgement (CJ), where subject experts judge two or more examples of candidate work against each other and these judgements are aggregated and analysed statistically, offers an opportunity to meet some of those challenges, but much is unclear about how to operationalise it and its credibility for stakeholders. OCR has conducted a series of trials of CJ methods and software solutions, using examiners as judges. We focused on trialling different kinds of subjects and different forms of CJ, and gathered data on the reliability and validity of the methods, judges' opinions of the task and the utility of CJ for awarding. Trials consistently produced judgements with high levels of internal reliability. Judges were capable of making holistic judgements in a meaningful, timely way, though some had difficulty with the fact that judging is different psychologically from the more familiar task of marking. A change to the Bradley-Terry model-based approach to analysing judgement data, called the Simplified approach, helps improve its efficiency and timeliness. Implications of the use of CJ outcomes in awarding in terms of procedures, and for standard maintenance and the credibility of grade boundaries, will be discussed.

What is the exact assignment for the judges to base their judgement on?

We ask judges to make a holistic judgement of the scripts they are looking at and to determine which is better. We suggest that decisions should be made relatively quickly, and that gut reaction is acceptable. Judges can use whatever method they wish to determine the quality of the scripts, but are asked not to re-mark the scripts. The exact question asked of them within the Comparative Judgement tool has changed over time. Currently it is "Which shows better overall performance?", but we have also used "Which candidate has performed better?" Ofqual use the question "Which of these two scripts is better, based on a holistic judgement of overall quality?" Judges are almost always markers and this, we think, will influence what construct they use to judge the scripts. Research into judgemental processes in comparative judgement is ongoing at OCR and within Cambridge Assessment.

How does this approach work when there is a specification change and one year's test/exam is in a different style and/or has different content and demands?

In some contexts, Comparative Judgement can put quite a heavy cognitive load on judges. In Comparative Judgement for awarding judges need to keep in mind the difficulty of the two tests being compared, and take that into account when deciding which piece of work is better. The more different the two tests being compared the more demand is put on judges. Fundamental to Comparative Judgement is that the construct on which the work is being judged (e.g., 'quality of work' or 'overall performance') can be applied to all work being judged. It would be asking a lot of judges to have them make judgements on scripts from tests with very different styles, content and demands.

What is the spectrum of stakeholder appetite for greater reliability, reduced bias in marking with Comparative Judgement, versus greater examiner remuneration?

*We have focused hitherto on Comparative Judgement as a replacement for *awarding*, not marking, so we are only talking about remunerating a handful of examiners per component, rather than the potentially hundreds used in marking. Nonetheless, it is true that, as Comparative*

Judgement costs about twice as much as traditional judgement in awarding (using the Simplified Pairs method) there is a difficult balance to strike here, and we are in continual consultation with internal stakeholders.

Did you deploy any selection criteria for recruiting judges or are they randomly selected examiners?

We use the people who are in charge of the standard of our different papers to identify the strongest judges in terms of their marking so these tended to be Principle Examiners or Senior Team Leaders, Team Leaders with a very good 'grade' in terms of their marking quality over the last however long they have been judging. We have plans in the future perhaps to do more work using slightly less experienced assessors but generally these were the best assessors we could find. That's party because they would generally be the most senior and the most experienced.

For awarding General Qualifications, the evidence of how much the papers differ overall is much less use than the different between papers at the judgmental grades. Have you used Comparative Judgement to look at this?

The Comparative Judgement procedure we discussed allows us to derive both a figure for the difference in difficulty between the two tests as a whole, and to equate each mark point across the score range. In our research about how Comparative Judgement outputs would be used operationally within awarding decisions, opinions have varied as to the most appropriate of these metrics; both are likely to be useful in different contexts. We intend to present decision-makers with both sets of information, and to continue to work to explore issues about which is the most useful. For a further discussion of this question, see Benton, Cunningham, Hughes and Leech (2020).

A concurrent item calibration method to improve measurement and support awarding of qualifications with optional items and papers: the example of GCSE History

Elena Mariani, Pearson

The debate around Examinee-Selected-Items (ESI) and optionality of papers in UK qualifications is longstanding yet still ongoing motivated by the fact that the existence of ESI is a threat to the measurement validity of an assessment (e.g., Morrison 1972; Bell 1997; Wang et al. 1995; Bramley & Crisp 2019). To guarantee equivalence of outcomes when there is paper choice, different cut offs are set in awarding according to paper combinations based on expert judgement informed by item and paper level statistics and driven by the comparable outcome approach. No statistical approach has been agreed inter-board to formally test the equivalence of the sets of boundaries established in this way. Similarly, although several methods exist to determine the comparability of ESIs (e.g., Liu & Wang, 2017; Bramley & Crisp 2019), there is no inter-board agreed method to adjust the final score for possible imbalances between alternative optional questions and the common practice is to treat alternative optional questions as equivalent.

In this paper we suggest a statistical model based on concurrent item calibrations with a Partial Credit Model to produce a qualification outcome measure accounting for ESIs and paper optionality. We also suggest a statistical procedure to link the outcome measure to the national prediction matrix to support awarding. We illustrate the model using the case of Pearson GCSE History; this qualification is made up of 27 distinct papers which can be combined in 240 possible routes as combinations of three components. Each candidate sitting a combination of three components encounters two ESIs. When all papers are considered all together, they form a network of items, where items from common papers taken by candidates on different routes become internal anchors for establishing equivalence of routes and ESIs.

For awarding General Qualifications, the evidence of how much the papers differ overall is much less use than the different between papers at the judgmental grades. Have you used Comparative Judgement to look at this?

Thank you for the suggestion. I agree that for the purpose of awarding the key factor is the difference at key awarding boundaries, while in my analysis I have assumed that the overall difference is the same at all boundaries. I have not looked at Comparative Judgement to look at this specifically. One challenge I see in running a Comparative Judgement with the history specification is that for many components we have limited entries, so it may be potentially challenging to obtain a sufficient number of scripts around the boundary marks. But definitely worth exploring.

On the topic of curriculum/question options. Curriculum options are in curricula for several reasons including local knowledge and upholding diversity and facilitating learning. Often Ministries are keen to include local knowledge rather than have an entirely "scripted" global curriculum. Also we know that what is assessed influences what is taught. Does this influence the importance of curricula and or question options/choice?

I agree that one of the advantages of optionality is the increased flexibility it offers to schools and the personalisation of teaching and learning experience. Unfortunately, there can be unintended consequences, as some schools may be "better" at choosing an option that suits their learners than others, with the consequence that learners may be at an advantage. Perhaps the decision on how much optionality to have in an assessment should be the result of a trade-off between the benefits that it allows and the risks of unintended disadvantage introduced by optionality.

The risk can potentially be mitigated by ensuring that centres and learners have full information when making curriculum decisions and awarding bodies have the responsibility to ensure that optional items are as equivalent as possible (e.g., language is accessible in the same way). I think with the work I have presented I wanted to highlight that there is still room for improvement in terms of mitigating the risks associated with optionality.

Year-on-year standard maintenance for small entry subjects

Liz Harrison, AQA

This presentation will consider the problem of maintaining year-on-year grading standards in small entry awards. Statistical information may be very limited and common practice is to carry-forward the previous year's grade boundaries as a starting point in awarding meetings, this assumes that the papers are of equivalent demand. Examiners can then adjust these boundaries on scrutiny but are likely to be anchored to the prior information. Alternative starting points could be offered by assuming that the ability of the students is equivalent and then use equating methods to link the two years. Resampling techniques were used to explore the performance of several equating methods using data from some well understood large-entry awards. Livingstone's circle-arc equating was found to be a good method. Circle-arc equating was then trialled with some success in a few small entry international awards last summer and we are looking to use it further in future.

I have a particular interest in Further Maths. Are you intending to re-run the study with very skewed distributions?

I agree it would be good to look at Further Maths–the mean would be high and the distribution skew. At the moment I do not have immediate plans to look at data like this and assess how circle-arc equating performs. But I hope to be able to do this in the future.

Session 5 - Maintaining Standards, part 2

(Chair: Tom Bramley)

Watch the session.

Using ISAWG to award GCSEs

Darren Johns, WJEC

At the Ofqual Educational Seminar in 2019, we used our first year of data to show that ISAWG, as a measure, stabilises relatively early during the awarding process. Now that we have two years of data, we have been able to use 2018 data to create base matrices and have simulated how ISAWG predictions would have worked in practice.

This research provides a comparison of ISAWG-based predictions with predictions used for GCSE awards using data from the summer 2019 awarding series. The analysis found that slightly different initial recommendations could have been expected if boundaries had been set based upon an ISAWG prediction and, that whether the ISAWG predictions suggested we awarded leniently or harshly varied from subject to subject.

Further, the analysis looks at the stability of the relationship between ISAWG and subject outcomes in different years, and will examine circumstances where ISAWG may or may not be viable as a major piece of evidence at awarding.

Benton, T., 2017. Pooling the totality of our data resources to maintain standards in the face of changing cohorts. AEA-Europe Conference.

How valid is that assumption, that deciles of ISAWG perform the same in different years?

The important thing is that in Wales WJEC have just about the full cohort (excluding those who don't take any GCSEs and those in independent centres taking all GCSEs with other boards). If we believe in comparable outcomes then there is no reason why the deciles would differ from year to year.

In England we would have to get a full cohort (i.e., at least one mark from every candidate doing any GCSEs). It is possible that changing popularity of subjects could change the meaning of ISAWG over time, but this is the same issue as mean GCSE and Key Stage 2 deciles.

What parameters would you use to determine the base year for this approach? Given the outcomes of summer 2020, would you use this data? What approach would you take for first awards (e.g., cognate subjects)?

These would be the same issues faced as when deciding the base year for existing prediction methods. There are other options, such as creating a base matrix built from combining multiple years of data, or perhaps just a selection of subjects. Usually setting the base year has been the responsibility of the regulator, and as a rule of thumb is set as the first two years after reform. We could not use summer 2020 data as the ISAWG relies upon paper scores which do not exist for this series.

Was the ISAWG score standardised between years on a broader definition of Stable Common Centre or Common Centre? Or do you assume that general ability is the same in both years?

As per the comparable outcomes approach we would assume the general ability of the cohort in both years to be the same. As we have the whole cohort in Wales this is very similar to a Common Centres prediction though it does include centres taking a subject for the first time who would be dropped from Common Centres.

Would restricting the ISAWG population to Stable Common Centre or Common Centre to account for the effects in say in drama/Welsh etc.? And yet still strengthen Stable Common Centre/Common Centre?

The issue with Welsh second language is that in the base year those who are good at Welsh second language took the GCSE and weaker candidates took the Short Course (i.e., relative weakness for their ability). Policy incentivised centres to put all candidates through the full course in 2019, so the relationship between ISAWG and the subject has changed quite a bit. This is akin to a change in the value added relationship which could be seen with more traditional predictions.

Since centres were mainly expanding or reducing their entry, Common Centre would probably not have had much of an effect. Limiting the prediction to Stable Common Centre however would probably have had the desired results with predictions likely to be very similar to a conventional Stable Common Centre prediction based on the same base year.

Deconstructing the sawtooth effect

Paul Newton, Ofqual

When qualifications are reformed, we tend to describe our expectations concerning candidates' examination performances (pre- versus post-reform) in terms of the Sawtooth Effect. We use the Sawtooth Effect as a justification for applying the Comparable Outcomes principle. But what is the Sawtooth Effect? Why does it occur? How does it occur? When does it occur? And for how long does it occur? We don't have very good answers to any of these questions. This presentation will attempt to deconstruct the Sawtooth Effect, to understand better its nature, its prevalence, and what to do about it.

To actually see a graph like the one you've put up, does that assume that you have some good statistical equating mechanism, say, like in the US, so you're pretty confident that you're actually measuring performance on the same scale from the ATA?

Yes, so I think it's important to emphasise that what we're looking at here is the hypothetical graph. It's really really hard to get under the skin of this to understand exactly what's going on. Most of the time we basically just assume there is some kind of 'Sawthooth Effect' going on, so some kind of disruption followed by an inauthentic rise over time, but we have very little hard and fast evidence of that. The best evidence that we've got so far comes from Bangov's study from 2016 and some assumptions went into that but they seem like plausible assumptions which helped to quantify what might be going on there in terms of superior's rises but again we've got very little insights into that. We kind of really need to get parallel independent measures of the construct that we are trying to assess in order to try and rule out these supposedly test-specific effects. I mean nowadays we've got the national reference test for English and Maths but even those effects, it's questionable whether they will be completely unaffected by some of what I've been talking about there, particularly as they are a bit more curriculum-made. So it's all a bit ambiguous I think still.

Do we need to think more about assessment design when each exam 'assessment is being written, rather than during awards?

So, for example, to try to avoid undue predictability? We have had a lot of debate over the past few years about predictability. If you can't predict what's coming up from one year to the next, then you're not going to get these Augmentation Effects occurring (i.e., coaching, reallocation). On the other hand, you don't want tests to be too unpredictable, such that, as a candidate you're just thrown when you get to the exam. Steve Holmes has done some work on this recently, and that's going to be published as well, soon; maybe alongside the work that I've been presenting today.

Any hope that the National Reference Test can help?

It might do, to some extent. However, to the extent that the national reference test is similar to our exams–and it is to some extent–it might well be affected by some of these inauthentic performance gains. It's hard to know–it's kind of a big thought experiment–which is why I quite like the idea of trying to get under the surface, analytically, to try to tease out the different facets of the thought experiment.

What is the Sawtooth Effect if the specification change aims to reduce the content required and/or increase the options?

Have a look through the main paper when it's published on the Ofqual website. I situate the

Sawtooth Effect within a broader conceptual territory, as a particular kind of Contestable Transition Impact. There are other Contestable Transitions Impacts, such as when the amount of curriculum time made available for a subject changes, which is similar to the idea of changing the amount of content that needs to be studied.

An investigation of the comparability of standards between routes to certification in GCSE History and English Literature, and A Level Sociology

Qingping He, Ofqual

Optionality, the use of optional questions in examinations where different candidates can answer different sets of questions to achieve the same total scores or grades, is wide spread in GCSE and A level qualifications. The comparability of grade standards between different routes to certification in qualifications involving optionality has always been a concern for relevant stakeholders. Ofqual is currently conducting research to try to understand how optionality in general qualifications works and, in particular, whether common standards are applied to different routes to certification within a qualification, with a view to improving its regulatory functions. As part of this research, the comparability of scores between optional questions and grades between optional routes at both component and the overall qualification levels for seven specifications from GCSE English Literature, GCSE History and A Level Sociology was assessed using a number of statistical methods. Main findings from this work will be presented.

Should we stop offering question choice?

In general, the use of optional questions in the exam papers (and the use of optional papers in some of the assessment components) in the subjects studied to a large extent reflects the nature and structure of their contents and how they are delivered in schools. In particular, they reflect curriculum choice such that teachers can select topics relevant or appropriate to their classes and/or which reflects their own expertise. Such subjects, as they are currently structured and delivered, depend upon the use of optional questions in their assessments. Although the data analysed are specific to these subjects, findings from this study do suggest that setting comparable performance or grade standards for a multitude of different routes to certification can represent a very challenging task for any qualifications. This should be taken into consideration when designing curriculum and assessments. If optional questions are to be used, appropriate and valid procedures should be in place to take account of differences in difficulty in optional questions when setting thresholds for performance standards to ensure that results are reliable, valid and comparable.

Some stakeholders have suggested increased optionality as one option for 2021. A bad idea?

Although the data analysed in this study are specific to the subjects investigated, the findings do suggest that setting comparable performance standards for different routes to certification would represent a very challenging task for any qualifications using optionality. This should be taken into consideration when designing assessments. The use of optional questions in exams itself is not an issue. The issue is how we can ensure that results are valid and fair for candidates taking different optional questions or optional papers. If optional questions are to be used, valid procedures should be in place to ensure that results are reliable, valid and comparable (in terms of performance standards) across different routes to certification and fairness is achieved. This appears to be particularly important as there is evidence in the research that less able students are more likely to choose more difficult questions. This underlines the importance of having valid procedures in place to ensure that grading does properly take into account differences in difficulty in optional questions.

If we stop offering optional questions how do we assess localised knowledge, local issues and include a diversity of knowledge?

There are probably two ways to cater for curriculum choice. One is optionality. The other is the use of more 'generic' questions which allow students to answer using different curriculum. Arguably, the latter is less satisfactory as the questions can be too vague and examiners are not always acquainted with the range of responses. If optional questions are to be used, valid procedures should be in place to ensure that results are reliable, valid and comparable (in terms of performance standards) across different routes to certification and fairness is achieved.

If better candidates choose the question where they're better able to demonstrate their knowledge and understanding, is it as simple as saying that that question is "easier"?

It is difficult to make a judgement about whether the "better" performance on a question (or a section or a paper) is because the student has better knowledge and understanding or because the questions is easy without a basis on which both the student and the question are compared at the same time. Even if more able students are more likely to select easier questions, this is not a problem as long as we have appropriate procedures in place to ensure that the final grades are comparable in terms of performance standards for all candidates.

Was there evidence of centre level effects in the difficulty of optional papers/items?

This is an interesting question, but we have not looked into this yet.