Education Assessment Research Seminar

12 and 13 October 2020

Online dissemination of work by leading UK education assessment researchers

Each presentation will be no longer than 20 minutes including time for questions.

As this is an online seminar we are taking the opportunity, with the permission of the presenters, of widening 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.

Monday 12 October 2020

Session 1 - Vocational Qualifications (Chair: Jackie Greatorex)

09:00 – 10:30 BST

1. ***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).

2. ***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.

3. *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.

4. *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

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.

5. 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.

Session 2 - Accessibility (Chair: Gill Elliott)

11:00 – 12:30 BST

1. ***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.

2. ***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.

3. ***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.

4. 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.

5. *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 AEs, 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.

Tuesday 13 October 2020

Session 3 - Marking (Chair: Irenka Suto)

09:00 – 10:00 BST

1. ***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.

2. ***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.

3. *'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.

Session 4 - Maintaining Standards, part 1 (Chair: Tom Bramley)

10:15 – 11:30 BST

1. ***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.

2. **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.

3. ***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.

4. *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.

Session 5 - Maintaining Standards, part 2 (Chair: Tom Bramley)

11:45 – 12:45 BST

1. ***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.

2. ***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.

3. ***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.