Centre for Education Research and Practice (CERP)
Impartial and rigorous research

Find out more at: cerp.org.uk
Centre for Education Research and Practice (CERP)
About AQA’s Centre for Education Research and Practice (CERP)

AQA’s Centre for Education Research and Practice – known as CERP – is a multi-disciplinary organisation that specialises in assessment research. Its work informs AQA’s operational activities and assessment design, and contributes to the wider debate on examinations and education.

CERP has its roots in AQA’s predecessor organisations, and is able to draw on the large body of research undertaken over the past decades. Much of this focuses on setting and maintaining standards, which continues to be a crucial area of research. However, our interests have developed and we also explore topics such as validity, marking and accessibility of assessments. Although our work is grounded in the practical realities of qualifications, we are also interested in theoretical and philosophical thinking about assessment and qualifications.

Collaborating with colleagues across the assessment research spectrum is of great importance to CERP, and our researchers regularly undertake joint projects with partners such as universities and other research units.

AEA-Europe’s annual conference is a vital place for those of us within the assessment community to share ideas and challenge ourselves to think critically about our own research. This year’s theme – ‘Social and political underpinnings of educational assessment: past, present and future’ – has inspired our researchers to delve into a broad selection of subjects; abstracts for all the work that CERP will be presenting in Limassol can be found within this booklet.

We look forward to meeting colleagues new and old. CERP will have an exhibition stand in the refreshment area throughout the duration of the event; please do pay us a visit.

Alex Scharaschkin
Director, CERP
Improving students’ future prospects or extending the reach of the accountability framework? Investigating the impact of the English Baccalaureate on the educational landscape

Emma Armitage
Presentation

In England, students are assessed around age 16 via the General Certificate of Secondary Education (GCSE).

The outcomes of GCSEs serve as indicators of both student and school performance. Consequently, conflicts have arisen between assessment that serves student interests and assessment for accountability. The English Baccalaureate (EBacc) – a performance measure that reports the percentage of students who take GCSEs in English, maths, science, history or geography and a foreign language – provides an interesting example of this tension.

Although the EBacc is not yet a formal accountability measure, there are plans to require 90 per cent of students in future cohorts to enter for it. Education ministers argue that studying these rigorous academic subjects will keep students’ options open for the future. However, critics contend that this may be detrimental to the attainment of students who are not academically inclined, or whose interests lie elsewhere.

The current study analyses how EBacc uptake and attainment has changed between 2010 and 2014, taking into account prior attainment and school type. The findings will be discussed with reference to the interaction between policy and practice, in particular the potential ramifications of entering 90 per cent of students into the EBacc alongside the introduction of Progress 8, which will measure pupils’ progress across eight subjects from age 11 to 16.
A site of tension: The complex case of GCSE English speaking and listening

Ruth Johnson
Poster

Research suggests that high-stakes testing leads to a narrowing of the curriculum, as teachers focus on teaching to the test. In England, the GCSE in English is a case in point. Analysis of this qualification sheds further light on the tensions between teaching and learning, and the priorities occasioned by accountability measures.

Qualitative studies were undertaken to explore the response of students, teachers and examiners to the GCSE in English. This poster focuses on the speaking and listening component, which constitutes 20 per cent of the overall assessment. A total of 23 semi-structured interviews were conducted with students and teachers across three schools, and with senior examiners.

The study found that the senior examiners conceptualised the speaking and listening assessment idealistically, viewing it as a site of empowerment for students from all social and cultural contexts.

However, students and teachers tended to conceptualise speaking and listening as discrete assessment tasks. In very few cases was speaking and listening understood as a site of teaching or learning.

D-optimal adaptive comparative judgement

Yaw Bimpeh
Presentation

Comparative judgement has been applied to a variety of educational purposes in recent years. It offers an alternative to marking procedures, and may be preferable in certain contexts. However, due to the large number of possible script pairs, full pair-wise comparison is not always feasible.

The existing adaptive algorithms for comparative judgement are often ad hoc with little or no formal basis. In order to reinforce fairness in educational assessment, it is important that there is a sound basis for any adaptive pairing of scripts.
This paper discusses the issue of choosing between the alternative options in comparative judgement. We analyse comparative judgement under the Bradley-Terry model and the proposed adaptive optimal design for paired comparison. The theory of experiment design is applied to provide efficient selection of sets of pairs for comparative judgement. The cyclic incomplete block design and general equivalence theorem are adapted to the Bradley-Terry model to provide an effective method for constructing and checking D-optimal comparative judgement.

We compare the performance of our method on simulated data against full factorial paired comparison. The results show similar patterns concerning perceived quality of the script for both methods. However, the proposed adaptive comparative judgement method significantly reduces the number of comparisons needed for full comparison, and is more efficient.

Improving the maintenance of standards in England: evaluating a comparative judgement approach to awarding
Kate Kelly, Charlotte Stephenson, Neil Stringer and Faith Jones
Presentation

To maintain standards over time, students of the same subject ability should receive the same grades whenever they sit their examinations. In England, grade boundaries are adjusted to compensate for vagaries in paper difficulty. However, current procedures cannot adequately monitor changes in difficulty.

Comparative judgement (CJ) presents a potential solution, as it allows paper difficulty to be evaluated directly and robustly.

In theory, differences in difficulty can be quantified through including items from multiple papers in the judgement exercise. These differences in difficulty can then be linked to changes in grade boundaries using item-facility test equating. This presentation reports on a study to evaluate this approach in practice.

The study aims to equate two AQA A-level Chemistry papers, using CJ difficulty estimates.
Standard setting/maintaining and public trust in national examinations around the world

Lena Gray, AQA; Tina Isaacs, Institute of Education; Jo-Anne Baird, Oxford University Department of Education; Dennis Opposs, Ofqual; Christina Wikström, Umeå University, Sweden; and Anton Béguin, Cito, the Netherlands

Discussion group

This discussion group will present findings from phase one of a project that aims to critically examine policy positions and processes for assessment standards in a range of countries. Presenters will briefly outline their country’s policy positions on standards, indicate the sorts of processes involved in setting and maintaining those standards, and highlight the key issues in public debates about standards in their country. The case studies will cover England, Sweden and the Netherlands.

Discussion about the extent to which initial findings contrast with, or coincide with, views of assessment standards in other countries will help to move the project forward. The discussion will illuminate similarities and differences in conceptual bases, operational approaches, and outcomes for students in participants’ own contexts, as well as in the cases presented. It is anticipated that the discussion will involve challenge to current theory on standards, as well as critical reflection on how national organisations approach standard setting and maintaining.

The discussion group will be of interest to researchers, policy makers and practitioners interested in assessment standards.
Assessing individual participation in collaborative group work
Ayesha Ahmed, University of Cambridge; Ruth Johnson, AQA
Presentation

Collaborative problem-solving skills are critical for students to prepare for the world of work. They are also powerful tools for learning. It is important that students from all social backgrounds are given the opportunity to learn and develop these skills. However, this is not current practice in many classrooms in England.

This presentation will share initial results of a study that trials methods for assessing 15 year olds as they work collaboratively to solve a computing problem. Our data includes teachers’ ratings of the quality of students’ discussions, students’ problem-solving processes and their solutions, teachers’ comparative judgements of performances, and students’ self and peer assessments. Detailed analysis of group discussions, using established coding schemes, allows us to identify solution-critical interactions and the utterances leading to these. This, in turn, allows us to form conclusions about the interaction between discussion skills and the quality of solutions, which informs our understanding of the construct and how it should be assessed.
Exploring the effects of undertaking the Extended Project Qualification (EPQ)
Charlotte Stephenson
Presentation

Current policy in England restricts the amount of school-based assessment within general qualifications. Project-based learning (PBL) can enhance students’ academic performance relative to traditional content-based teaching methods by increasing engagement, self-direction and motivation. Research into PBL usually explores its effects on academic performance within the same discipline as the project undertaken.

However, recent evidence suggests that students working towards the Extended Project Qualification (EPQ) – an optional qualification offered in addition to GCSEs and A-levels – showed enhanced academic performance in other subjects.

The research presented here contributes to the extant literature by exploring teachers’ perceptions of the effects of the EPQ on students’ performance in other academic disciplines, and their academic performance generally. A qualitative investigation explored the experiences of 20 EPQ teachers in focus groups. Thematic analysis was used to analyse the data and identify common themes. The findings elucidate the potential benefits of PBL and inform how this link could be investigated quantitatively.

This paper also reveals teachers’ perceptions of the effects of school-based assessment at a time when examinations are the preferred approach to educational assessment in England.
Do different social groups experience differential item functioning (DIF) on items with certain features in GCSE English and GCSE Mathematics?

Ben Smith and Ruth Johnson

Presentation

The General Certificate of Secondary Education (GCSE) is a high-stakes qualification that is generally undertaken by students aged 16.

This paper reports on a research project that examines GCSE English and GCSE Mathematics. These qualifications are analysed at item level to identify items, or types of items, that appear to advantage or disadvantage groups of students.

Items and mark schemes were coded using a number of parallel typologies, which considered the different dimensions of the items. This allowed the researchers to consider whether trends in differential item functioning (DIF) emerged across items with certain features.

The effect of a range of contextual and demographic factors (including gender, ethnic group and socio-economic status) on item responses was analysed using latent class logistic regression methodology. This allowed for the possibility that not every member of a certain social group will exhibit the specific traits that make them more or less likely to perform well on a given assessment, and thus resulted in a more nuanced analysis of DIF than is typically undertaken.

Inter-subject comparability: how does adjusting grade boundaries affect schools, subjects and candidates in England?

Caroline Lau, Simon Eason, Ben Jones and Mike Cresswell

Presentation

In the national examination system in England, inter-subject comparability is problematic, both from a philosophical and practical point of view. There is an ongoing debate over whether some subjects are ‘harder’ or ‘easier’ than others. The extensive literature on this matter tends to focus on how to ‘correct’ the
perceived misalignment of subjects rather than on the practical outworkings, effects and implications of this.

This paper looks at the practical effects of aligning subject standards by using comprehensive, national GCSE results data from England in 2015. Subject grade boundaries were adjusted according to their relative grade difficulty, calculated using a Rasch analysis, and students’ marks were regraded. Following this, a range of pre- and post-adjustment attainment measures were calculated and compared to see how bringing the subject standards ‘in line’ would affect individual students, subjects and schools.

At subject level, it was observed that aligning standards has little impact on the proportions of students obtaining the same or similar grades in any pair of subjects. For schools and students, the changes in grade boundaries had little or no effect on their rankings. This would indicate that the complex task of aligning subjects statistically, even if philosophically defensible, may not be a worthwhile effort.

Developing constructed response (CR) test items
Ezekiel Sweiry
Workshop

The purpose of this workshop is to present and discuss guidance on developing constructed response (CR) items and their mark schemes. The guidance is based on a synthesis of available research literature on CR item writing, relevant aspects of cognitive psychology, and the presenter’s own experience of high-stakes test development across primary and secondary education in the UK.

A range of CR item-writing issues will be explored, including the appropriate use of language, real-world contexts and diagrams. We will also consider the features of items, mark schemes and examinee responses that can affect marking reliability, and, ultimately, how mark schemes can be designed to maximise marking reliability. Mark scheme validity will also be explored. Finally, the workshop will consider how qualitative and quantitative evidence
from item trialling can be used to identify problematic items and mark schemes. There will be practical activities and example test questions, and participants will have the opportunity to review potential revisions to a variety of sample questions and mark schemes.

The focus of this workshop will be primarily on short CR items and their mark schemes. The main emphasis will be on ensuring that items and mark schemes are, as far as possible, free of construct irrelevant variance.

---

**Applying formal concept analysis in assessment: can it help mediate between socio-political and technical understandings of the meaning of exam grades?**

*Alex Scharaschkin*

Presentation

UK GCSE and A-level examinations are curriculum-embedded assessments with a strong emphasis on tasks that require extended, constructed responses, such as essays, drawings and spoken presentations. Performance standards for these assessments are socially constructed and negotiated.

Fairness entails that a ‘grade A in English’, for example, should be equivalent across awarding organisations and over time. This would seem to require a constant, definitive account of the meaning of a ‘grade A in English’. But, can this be reconciled with a qualitative, evolving, and necessarily fuzzy account of what a grade A performance ‘looks like’, which reflects professional understanding of the assessment objectives?

This paper will argue that models that have been applied to analyse meaning in fields such as formal semantics and knowledge representation can play a part in bridging the gap.

It will examine the potential application to UK-style public examination assessment of formal concept analysis – a mathematical approach to deriving a concept hierarchy from a collection of objects and their properties – and its use in facilitating dialogue between subject experts, assessment experts, and policy makers.
The CERP team at AEA-Europe

Emma Armitage
Emma joined CERP in May 2015, having completed a PhD in Psychology at Lancaster University. She also holds a BSc in Psychology and an MSc in Developmental Disorders. Emma’s doctoral research explored three- to eight-year-olds’ understanding of pictures as symbols, with a specific focus on emerging knowledge of different picture mediums. Her paper on this topic was published in the journal of Developmental Psychology. Emma has taught on various undergraduate modules at the universities of Lancaster, Salford, and Central Lancashire.

Yaw Bimpeh
Yaw joined CERP in September 2014. He holds a PhD in Statistics, an MSc in Mathematical Sciences and a BSc (Hons) in Mathematics. His current areas of research include marking reliability, optimal adaptive comparative judgement, application of the Bayesian method to standard setting, and methods for detecting exam security breaches. Yaw has experience of analysing and modelling data in a variety of fields, and is skilled in the research and application of statistical methods. He has also taught statistics and mathematics to undergraduate students.

Lena Gray
Lena joined CERP as Head of Research in July 2014. Prior to this, she was Head of Service, Policy, Assessment, Statistics and Standards at the Scottish Qualifications Authority (SQA). Lena oversaw the SQA’s programme of monitoring standards over time, and was responsible for quality assurance policies. She also has experience as a teacher in secondary schools and as a tutor at the University of Strathclyde. Lena is currently working on a major international project investigating standard-setting approaches in a range of jurisdictions around the world.
Ruth Johnson
Ruth joined CERP in June 2015, having spent five years in the AQA English team. She has 15 years’ experience as a secondary English teacher, including five years as an assistant headteacher. Ruth completed a Doctorate in Education (EdD) at the University of Manchester’s Institute of Education and has a BA (Hons) in English from the University of Cambridge. She has a particular interest in the relationships between policy, assessment and school-based practices.

Kate Kelly
Kate joined the team in June 2010 after completing a BSc (Hons) in Psychology at the University of Bath, which included a year spent with CERP as a placement student. Kate’s specialism is comparative judgement; at last year’s AEA-Europe conference in Glasgow she presented a poster on a novel method of test equating using comparative judgement difficulty estimates and item facilities. She is currently working towards a PhD on the potential of comparative judgement for improving grading decisions.

Caroline Lau
Caroline joined CERP in October 2015, after completing a postdoctoral position modelling biological systems. She has a PhD in Systems Biology and a BSc in Computational Biology, both from the University of Manchester. Her knowledge of processing and analysing large datasets is used to improve the way CERP approaches data and computing tasks. Caroline’s current research focuses on how inter-subject comparability may affect school accountability measures. She also provides technical and statistical support during AQA’s awarding process.
Alex Scharaschkin
Alex became Director of CERP in 2014, having been a member of CERP’s advisory group for four years. He was previously Director for Regulation, Consumers and Competition at the National Audit Office (NAO) in London, where he led the NAO’s work examining the government’s use of markets in the private and public sectors. Alex has a background in assessment research: he was Principal Officer for Statistical Analysis at the Qualifications and Curriculum Authority, and held research posts at the Associated Examining Board and at London University’s Institute of Education.

Ben Smith
Ben joined CERP in September 2014, shortly after completing an MSci in Psychology and Psychological Research at the University of Birmingham. He is currently working with Yaw Bimpeh to develop measures of marking reliability for AQA’s assessments, and exploring a Bayesian method of standard setting. Ben is also interested in fairness of assessment and is working with Ruth Johnson to investigate this at individual item level. At last year’s AEA-Europe conference, Ben presented a methodological paper that detailed spatial analyses of educational outcomes.

Neil Stringer
Neil joined CERP in April 2005 from the Department of Psychology at the University of Surrey, where he was a postdoctoral Research Fellow in the field of applied human vision. As Principal Research Manager, Neil has responsibility for research in the area of assessment quality. His own current work is a mixture of operational research as well as broader research and policy work, for example on standard-setting procedures and admissions to higher education.
Charlotte Stephenson
Charlotte joined CERP in 2014. She holds a BSc (Hons) in Psychology from the University of Manchester. Charlotte’s research activities have included a study into the effects of enhanced team leader feedback on marking reliability and examiner satisfaction, and an investigation into whether comparative judgement estimates of question difficulty can be used to set grade boundaries for GCSE and A-level examinations. She is currently exploring students’ and teachers’ perceptions of the effects of undertaking the Extended Project Qualification on students’ academic performance.

Ezekiel Sweiry
Ezekiel joined CERP in July 2015. He has 16 years’ experience in test development and assessment research, and has worked for the Department for Education and leading assessment organisations. As a test developer, he has been involved in the development of a range of high-stakes tests in England at both primary and secondary level. His particular research interests include the factors that affect the difficulty and accessibility of test items, the item and mark scheme features that affect marking reliability, and the comparability of paper-based and computer-based assessments.

Find out more at: cerp.org.uk
Centre for Education Research and Practice

Edited by Claire Jackson and Lindsay Simmonds