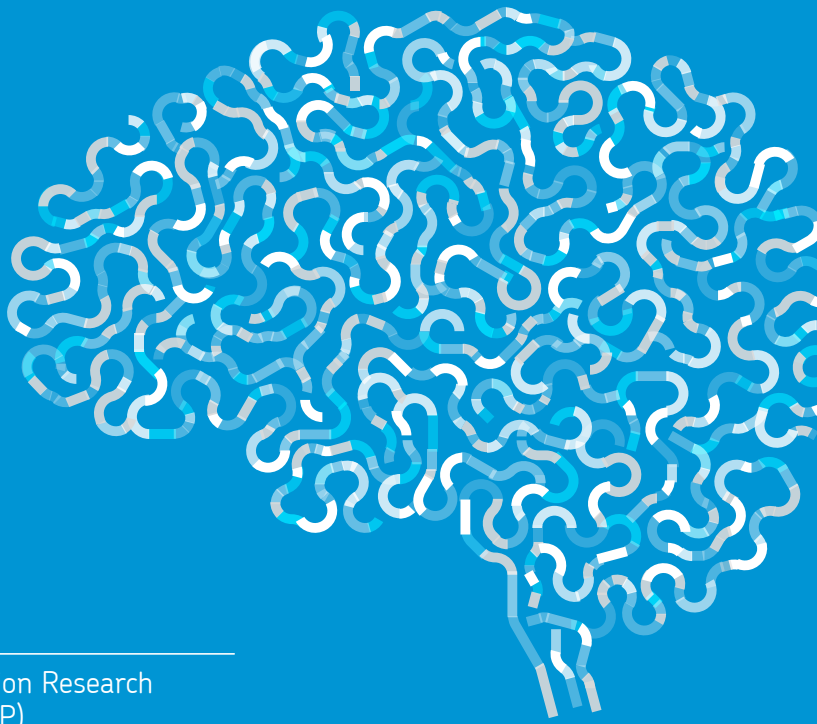
AQA
Realising potential

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 (CERP) is a multidisciplinary unit that specialises in assessment research. Its rigorous analysis and high-quality academic research inform AQA's operational activities and assessment design, and contribute to the wider debate on examinations and education.

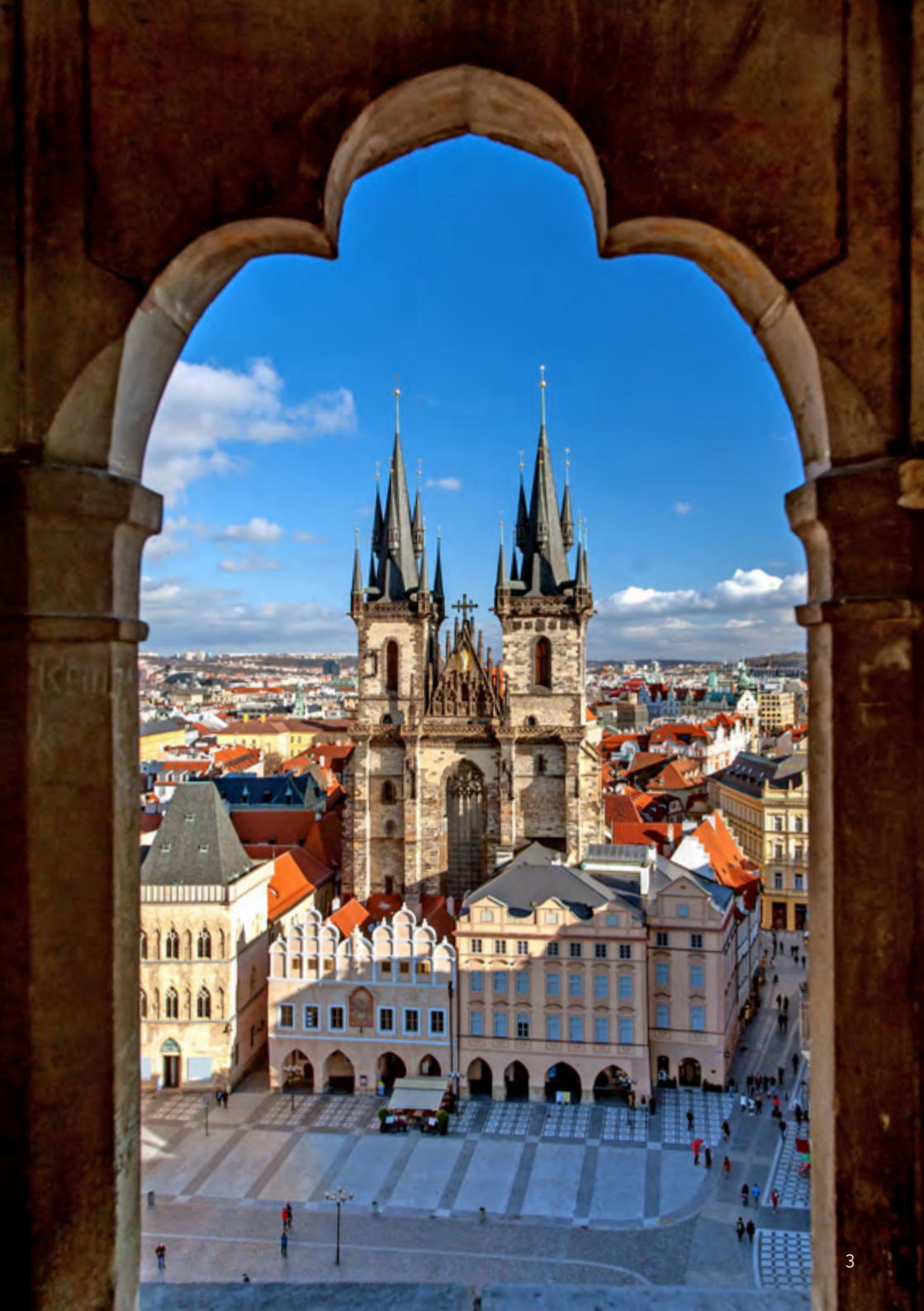
Much of our work focuses on improving existing methods of setting and maintaining standards. However, we also explore topics such as validity, marking reliability and the accessibility of assessments. Although our work is grounded in the practical realities of qualifications, we take an interest in theoretical and philosophical thinking about assessment and qualifications.

CERP's researchers regularly undertake joint projects, collaborating with partners such as universities and other research units. We also actively engage with the education community at large.

AEA-Europe's annual conference is a key opportunity for those of us within the assessment community to share ideas and critically challenge our own thinking. This year's theme – 'Assessment cultures in a globalised world' – has inspired our researchers to delve into a wide selection of assessment matters; abstracts for all the work that CERP will be presenting in Prague can be found within this booklet.

CERP will have an exhibition stand in the refreshment area throughout the duration of the event; please do pay us a visit.

Lena Gray
Director of Research



Exploring students' experiences of the Extended Project Qualification

Charlotte Stephenson

Presentation

Current policy in England restricts the amount of non-exam 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 post-16 qualification offered in addition to A-levels in the UK – showed enhanced performance in other subjects.

The research presented here contributes to the literature by exploring students' perceptions of the effects of the EPQ on academic performance. Significantly, it reveals students' perceptions of non-exam assessment in a nation where examinations dominate educational assessment.

A qualitative investigation, using semi-structured interviews, explored the experiences of 15 EPQ students. Thematic analysis was used to analyse the data and identify emergent themes. The students felt that the EPQ provides opportunities to build learner agency and self-awareness, advance academic skills, prepare learners for future success and improve engagement; hence it has the potential to fill gaps in the post-16 curriculum. However, these benefits are constrained by the perceived lack of recognition of the qualification.

Beyond classical statistics: different approaches to evaluating marking reliability

Ben Smith, **Elizabeth Harrison**, **William Pointer** and **Yaw Bimpeh**

Symposium

In high-stakes national examinations, quality of marking is a key concern; if students are not awarded the 'correct' grade or score, the integrity of both the examination system and any selection or monitoring based upon examination results is called into question.

In recent years, there has been a wealth of research into marking accuracy in England. Typically, this research has co-opted examiner monitoring data to assess the reliability of marking post hoc. It has generally adopted a classical approach to marking reliability – namely that the final mark a student is awarded equates to their 'true score' combined with an error component.

This symposium features three presentations that consider different approaches to quantifying marking reliability: Generalisability theory (G theory), the many-facet Rasch model and confirmatory factor analysis.

G theory splits the error into variance components, which can then be attributed to the different facets that may contribute to unreliability in marking (such as the marker, the question or the response).

The many-facet Rasch model extends the rating scale model and has the versatility to allow a number of different examiner biases to be explored.

Confirmatory factor analysis differentiates between true score and marker error, but crucially it resolves the confounding of random marking error and systematic error, allowing for model fit statistics to be estimated.

We discuss how the three approaches can be used to assess marking reliability, and explore the different information that each conveys about test items. We also evaluate each approach in terms of its practical value to test developers and researchers.

Giving G theory marker statistics a context: comparison to classical measures and post-marking data

Elizabeth Harrison

Presentation

In the UK, on-screen marking enables the quality of each marker's work to be continuously monitored. Examiners' marks for certain items are systematically checked against other examiners' marks for the same item.

G theory is one approach that can be used to estimate the scale and impact of marker error on students' marks and grades. This presentation discusses the post hoc application of G theory to marking monitoring data – the data includes instances of both poor and successful marking.

The results from the G theory analysis are compared to a classical approach to marking reliability. The comparison indicates that G theory can be used to produce equally understandable and useable statistics. These statistics also correlate well with the observed rate of mark changes from post-results marking reviews, which indicates that they give a reasonable prediction of the scale of marker error that affects all students.

Operational use of these statistics could facilitate a process of continuous improvement, enabling test developers to identify items associated with especially reliable or inconsistent marking and improve the design of items and mark schemes, so as to reduce marker error in the future.

Limitations of the data collection are discussed, together with possible improvements.

Putting a G theory approach to marking reliability through its paces

Ben Smith

Presentation

This presentation discusses the trialling and implementation of a novel approach to evaluating marking reliability based on G theory.

In on-screen marking systems currently used by AQA, the quality of examiners' marking can be monitored via two methods: seeding and double marking. Seeding requires large numbers of examiners to mark a handful of responses (seeds); double marking involves two examiners marking many responses that have been randomly selected.

During trials, it became clear that there were significant differences in the reliability values of items monitored using each of the two methods – possibly due to the different sampling techniques involved. It was also discovered that, in some cases, the spread of marks for seeded items was not representative of the spread of marks for all responses. This led to concern about the accuracy of candidate variance estimated using G theory methods.

The presentation explains how these two issues can be addressed – via a statistical manipulation to render the reliability statistics comparable for the two monitoring methods, and by using the variance for all responses to estimate candidate variance for seeded items.

Real-life examples are used to discuss the interpretation of G theory reliability statistics. There are several caveats that users of the method need to bear in mind, including small sample size and low candidate variance. Unrepresentative data is also a significant problem.

The presentation will outline how AQA is using these statistics as part of a continuous improvement process, and the training and resources that are needed.

What happens when extended response question papers are no longer divided into items for marking?

David West

Poster

This study looks at two different methods of on-screen marking currently used by AQA: item-level marking and whole-script marking. When marking is done at item level, each response is marked by a different examiner; in the case of whole-script marking, the entire paper is marked by one examiner.

In June 2015, 17 extended response question papers that had previously been marked at item level were switched to whole-script marking. This permitted a large-scale live empirical comparison between these two systems of on-screen marking.

More internal consistency was observed when the papers were marked as whole scripts by a single examiner: the average correlation between item marks rose by nearly 50 per cent. The spread of students' marks also increased when this marking approach was adopted.

In general, the marking was equally reliable in both years and correlation of marks with prior attainment of the candidates was unaffected by the change in marking approach.

Variance analysis appeared to support whole-script marking, indicating that a student's performance on the different items can account for most of the variation. However, this analysis was weakened by the very limited sample of items and scripts that had been marked more than once.

The change in marking approach had a greater effect on high-tariff item marks than on low-tariff item marks – of which there were only a minority contained in this study.

Evaluating the construct validity of educational assessment designs in the context of UK high-stakes qualifications

Yaw Bimpeh

Presentation

In many public examinations in England, and other parts of the UK, items within question papers are designed in accordance with assessment objectives (AOs). These AOs are set out in a regulatory framework by the government's Department for Education. AOs are hypothetical constructs that are not observed directly; they are measured indirectly through responses to items that are assumed to adequately represent the construct.

This study uses structural equation modelling to investigate the construct validity of the AOs in a test. The approach is applied to question papers from the new AS-level examinations (taken by 17 year olds) in Chemistry, Physics and Biology.

The adopted approach provides a framework to analyse AOs that are measured by multiple items. It uses an integrated statistical procedure that tests the measurement model and all related hypotheses at the same time.

With this model, we examine the properties of the measurement; for example, whether all the items are valid and reliable indicators for the assessment objective we want to measure. More importantly, this model enables us to answer questions about the validity of AOs, construct reliability and how different constructs are related.

We illustrate the method and discuss theoretical principles, practical issues, and pragmatic decisions to help evaluate the construct validity of high-stakes assessments in England.

Overcoming political and organisational barriers to international practitioner collaboration: guidelines for insider researchers working in exam boards and other public organisations

Lena Gray

Presentation

This presentation will outline issues for insider researchers working in exam boards and other public organisations concerned with high-stakes assessments, such as national school-leaving examinations. It will articulate political and organisational barriers to the research undertaken and suggest ways to overcome these.

Exam boards operate in a highly political environment, in which either the organisation or its individuals may be scapegoated for political failings. This can produce a risk-averse setting: while many exam boards may encourage reflection within the organisation, they may be uncomfortable with transparency outside the organisation. Exam board researchers can face enormous pressures in carrying out and disseminating research. This situation is detrimental to advancing understanding of theory, policy and technologies.

The presentation will suggest guidelines on how exam board researchers can critically analyse their personal and organisational practice and the dominant policy and cultural environment within their own national setting. It will explore how such researchers can be more transparent about the challenges they face.

The guidelines have been developed following a major project with practitioners in 11 countries: Australia, Chile, Hong Kong, England, France, Georgia, Ireland, South Africa, South Korea, Sweden and the USA.

The presentation will be of interest to researchers, policymakers and practitioners interested in transparency about assessment systems.

Is knowledge familiarity a good predictor of item difficulty? Rethinking Webb's (2007) Depth of Knowledge scale

Ezekiel Sweiry, AQA; **Yasmine El Masri**, University of Oxford

Presentation

Predicting item difficulty is a challenge that is relevant to all assessment practices. A large number of variables that influence item difficulty have been identified in the literature. Of particular interest is a variable we refer to as 'cognitive level', which typically ranges from knowledge or recall at the lowest level, to skills such as synthesis at the highest.

Cognitive level is often measured using Webb's Depth of Knowledge (DOK) scale. Previous studies have, in general, failed to show any relationship between cognitive level and item difficulty. This may be because items classified at the lowest level in DOK – recall – actually show considerable variation in difficulty; such variation may be due to differences in the familiarity of the knowledge assessed by these items.

This study, which uses science tests aimed at 11 year olds, investigates whether the use of separate rating scales for knowledge familiarity and higher-level skills would elicit a stronger relationship with difficulty than a single scale, such as DOK.

The paper explores key findings and implications of the study, including the success of the scales in terms of the variance in difficulty explained, the extent of inter-rater consistency achieved, and the challenges in constructing a scale designed specifically to address the familiarity of knowledge.

Analysing multidimensional ordinal data in attainment-referenced assessment

Alex Scharaschkin

Presentation

Assessment culture in England has resisted the imposition of largely closed-form or standardised testing models – such as the SAT in the US – with respect to national high-stakes examinations. Instead, the summative valuation of students' responses to tasks that require constructed responses (essays, performances, artwork, etc.) forms a substantial part of current assessment procedures.

UK national assessments are curriculum embedded, and it is necessary to demonstrate that students' overall results reflect the intended assessment objectives – that performances that are graded 'C', for instance, tend to exemplify the qualitative features that are supposed to be associated with 'a typical grade C performance'. In this regard, public examinations in the UK have been characterised as 'attainment referenced'.

This presentation seeks to model features (construct-relevant attributes) of performances in attainment-referenced assessment as mappings that associate with each performance an ordinal 'value'.

It will examine the possibility of using an analogue of principal component analysis for ordinal data to appraise the extent to which particular qualitative features are present in different classes of performances. Applications to the design of marking and grading procedures will be discussed.

Standard setting/maintaining and public trust in national examinations around the world: The effects of structural and contextual issues

Tina Isaacs, UCL Institute of Education; **Kristine Gorgen**, Oxford University Centre for Educational Assessment; **Lena Gray**, AQA; **Jo-Anne Baird**, Oxford University Department of Education; **Dennis Opposs**, Ofqual; **Iasonas Lamprianou**, University of Cyprus; **Anna Lind Pantzare** and **Christina Wikström**, Umeå University, Sweden; **Mary Richardson**, UCL Institute of Education; **Anton Béguin**, Cito, the Netherlands; **Nadir Zanini**, Ofqual

Discussion group

Standard-setting systems for curriculum-related, end-of-school examinations provide a powerful illustration of how well an educational system is doing, and are often the focus of debates about system fairness.

Educational cultures differ across jurisdictions, permeating assessment structures and processes in idiosyncratic ways. Hence the discussion of standard setting remains a bastion of the local in our globalised assessment world. A key question is: who has the power to set standards? Within any given national standard-setting system, the number, nature and status of bodies involved, and how they relate to each other, determine key features of that system. The way that the responsible bodies interact with wider stakeholders (such as examiners, teachers, parents and students), and how this changes over time, also has a major impact on the system.

This discussion group will include poster presentations that outline standard-setting structures in six national systems. A comparison of the cultural and contextual issues in each of the jurisdictions will provide a vehicle for exploring the effects of these issues on standard-setting systems. Participants will also be invited to discuss their own national system.

The discussion group will be of interest to researchers, policymakers and practitioners interested in assessment standards.

Assessing group dialogue: what is good participation in group work and how can we assess this?

Ayesha Ahmed, University of Cambridge; **Ruth Johnson**, AQA

Presentation

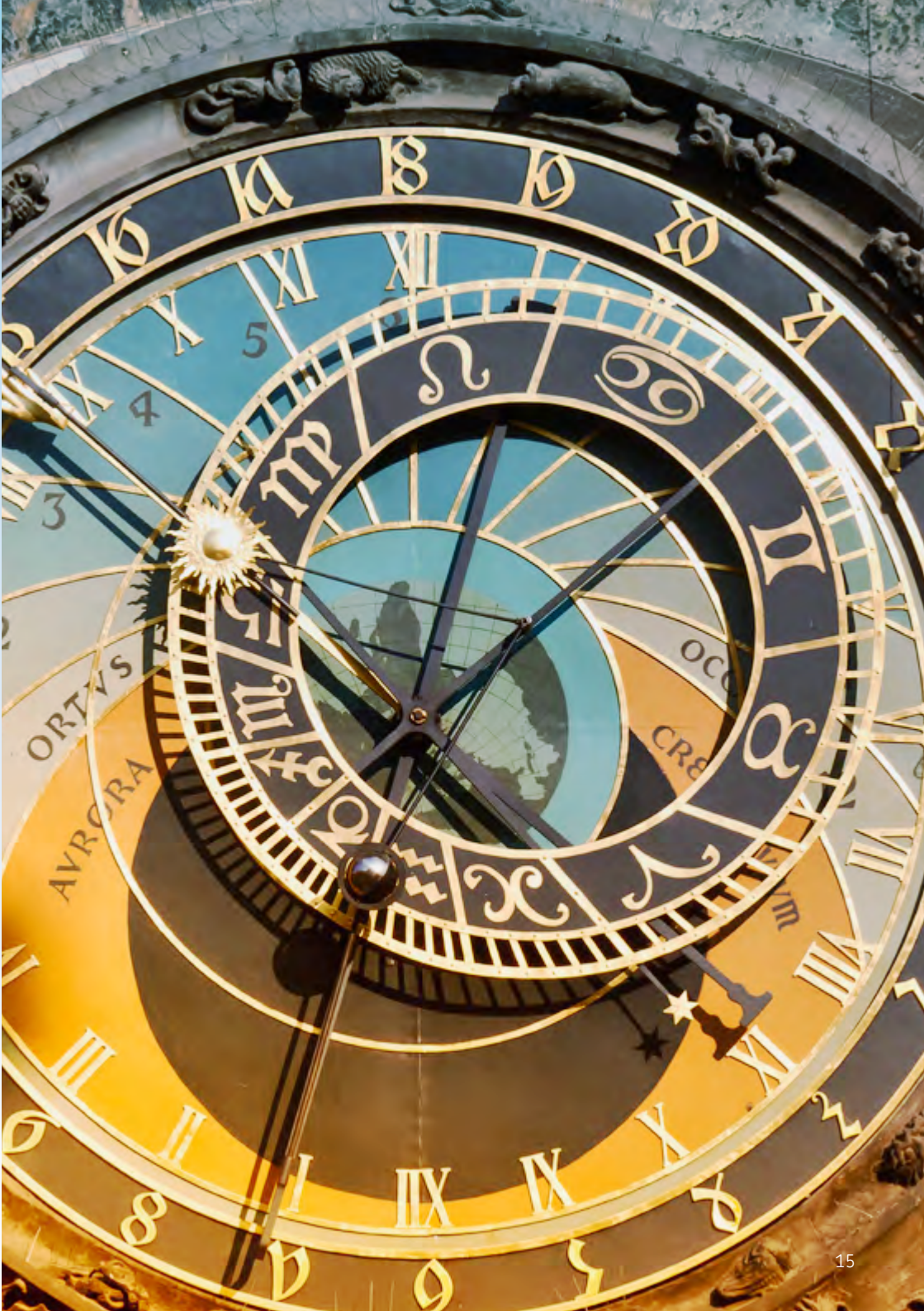
Collaborative skills have global prominence, as evidenced by the inclusion of collaborative tasks in PISA. It is therefore important to consider how to facilitate teaching and assessment of these skills.

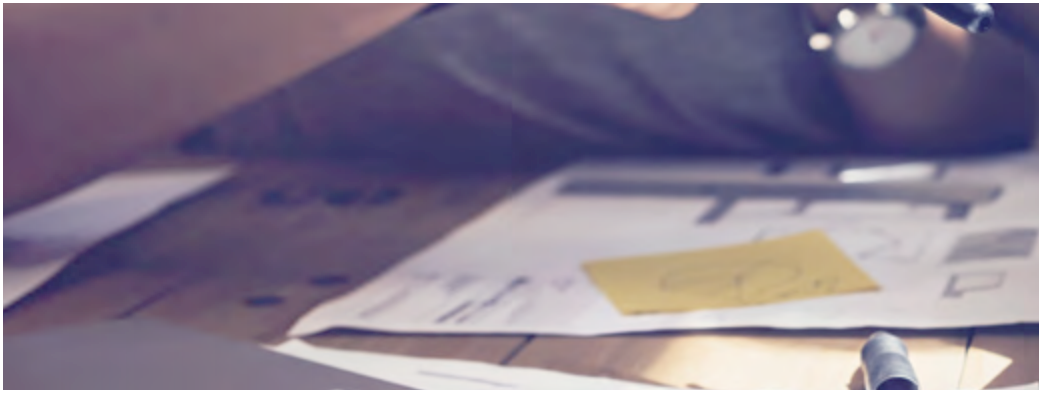
The political climate in England is not conducive to assessing collaboration: there is a lack of trust in teacher assessment, and school accountability is based on external value-added measures. Tensions between teacher assessments and external exams are particularly apparent in the assessment of skills that are critical for group work but less obviously crucial for written exams.

A better understanding of how to assess group processes may lead to collaborative skills becoming more valued in our curriculum.

We report on a study that investigates how features of group work can be assessed. We aim to identify features of dialogue that are important for good participation in group work, result in better outcomes of group processes, and can be assessed by teachers to inform teaching and provide useful feedback for learners.

The study focuses on 15-year-old students participating in robotics tasks. We filmed the group work, collected teachers' observational notes, and asked teachers to make comparative judgements of students' performances with regards to the discussion, problem solving and social elements of the interactions. Our approach to analysis is guided by a socio-cultural perspective in which solutions reached during the problem solving arise through the co-construction of meaning.


ORTVS
AVRORA



The CERP team at AEA-Europe

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, application of the Bayesian method to standard setting and test equating, and construct validity of assessment designs. 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 in July 2014, and was appointed Director of Research in May 2017. Previously, 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.

Elizabeth Harrison

Elizabeth joined CERP in September 2015. She previously worked as a data analysis manager at a secondary school, where she supported staff in target setting, monitoring progress, and understanding performance measures. Elizabeth also has experience as a research assistant at Nottingham University, and as a statistician in the pharmaceutical industry. She has a BSc in Mathematics with Statistics. Elizabeth's current research activities include supporting AQA's work on marking reliability and exploring the accuracy of vertical scaling in tiered papers.

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, and the assessment of skills that are difficult to measure.

William Pointer

William joined AQA in 2010, having graduated with a Masters in Mathematics from the University of Bath. He initially worked as a Qualification Developer in the GCSE Science department, before joining CERP in August 2013 to pursue a career in research. William's current research focuses on the quality of marking: how we measure and monitor marking reliability, and how we can improve it. He is also interested in research on standards and comparability.

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 colleagues on the development of a suite of marking reliability metrics for AQA's assessments. Ben is also interested in fairness of assessment and is working with Ruth Johnson to investigate differential item functioning (DIF) in GCSE and A-level question papers; preliminary findings were presented at last year's AEA-Europe conference.

Alex Scharaschkin

Alex became Director of CERP in July 2014, and was appointed Executive Director of Research and Compliance in November 2015. He was previously Director for Regulation, Consumers and Competition at the National Audit Office (NAO), 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 the Institute of Education, University College London. Alex also served as a member of CERP's advisory group for four years and is currently Executive Secretary of AEA-Europe.

Charlotte Stephenson

Charlotte joined CERP in 2014. She holds a BSc (Hons) in Psychology from the University of Manchester and has recently completed her Master of Research. 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 17 years' experience in test development and assessment research, and has worked for the Department for Education and leading UK 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.

David West

David joined CERP in June 2016, having worked in physics research for 18 years. He also spent six years working on secondary school data at a time when significant changes to curriculum and performance measures were introduced. David was a senior lecturer at the University of Manchester, where he directed postgraduate study in Modern Optics and worked on technology transfer and commercialisation of scientific research. He holds a BSc and PhD in Physics, both obtained from the University of Manchester. His current research includes the analysis of question paper functioning and marking systems.

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

Edited by Lindsay Simmonds and Kate Kelly

AQA Education (AQA) is a registered charity (registered charity number 1073334) and a company limited by guarantee registered in England and Wales (company number 3644723). Registered address: AQA, Devas Street, Manchester M15 6EX.
October 2017