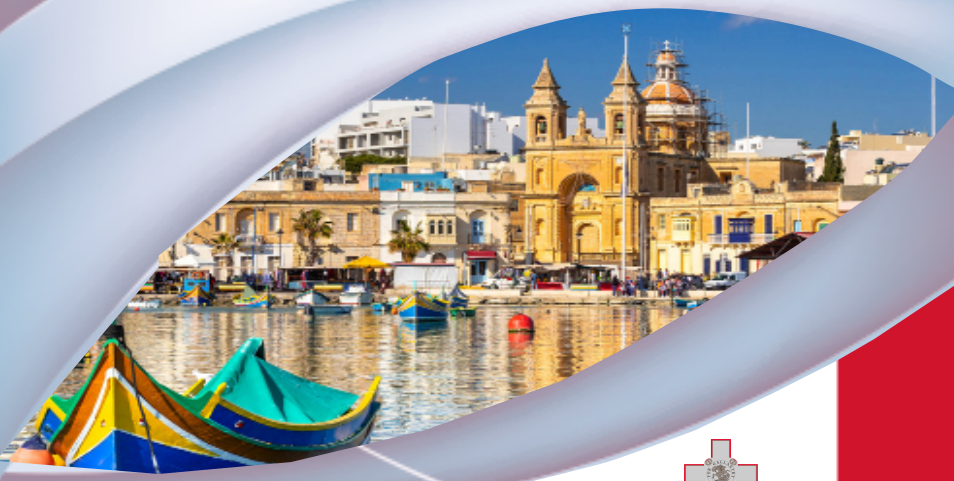
aea '23
EUROPE

01-04 November, 2023
Malta



# Assessment reform journeys:
## Intentions, Enactment and Evaluation

**BOOK OF ABSTRACTS**

# How hard can it be? The practice and challenge of validation: issues around how best to provide evidence for assessment validity, reliability and fairness

9:00 - 16:10

## How hard can it be? The practice and challenge of validation: Issues around how best to provide evidence for assessment validity, reliability and fairness

S. Shaw[1, 2]

[1]University of Cambridge, United Kingdom
[2]Institute of Education, University College London, United Kingdom

The primacy of validity as measurement concept has been consistently affirmed in the assessment literature. The responsibility for assessment providers to demonstrate robust and thorough validity evidence is, therefore, a long established expectation (Messick, 1992, p.89) as are warnings about the “potentially serious consequences” (Kane, 2009, p.61) of shirking such responsibilities. Even assessment providers that have limited resources will still have a responsibility to demonstrate the quality and validity of their assessments. This workshop is intended to make the complexities of validation theory and practice less challenging and more readily operational. The workshop broadly divides into two parts. The first establishes the importance and relevance of validation theory and practice. The second unpacks the complexities and challenges of conducting a validation exercise. Each part comprises sessions affording group activities and discussion. By sharing experiences through a collaborative workshop environment, greater insights will be drawn leading to an increased understanding of the validation process and how it might be routinely operationalised in differing contexts.

# Artificial Intelligence as a tool for Assessment Reform

9:00 - 16:10

## Artificial Intelligence as a tool for Assessment Reform

S. Bezzina[1], A. Dingli[1]

[1]University of Malta, Malta

In 2021, the Ministry for Education, Sport, Youth, Research and Innovation in Malta embarked on the EducationAI pilot project to introduce an AI-powered educational platform in Primary Mathematics. Part of the Government of Malta 'Strategy and Vision for Artificial Intelligence in Malta', the project aims to further move from a one-size-fits-all schooling system to a more equitable quality education for all. The project provides a context for implementing continuous assessment, recently introduced in the Maltese educational system as part of the Learning Outcomes Framework reform. Consequently, the workshop aims to enable participants to explore and discuss the potential of Artificial Intelligence (AI) in and for assessment reform through hands-on experience of the EducationAI educational platform. The target audience includes educators, administrators, policymakers, and researchers interested in exploring AI as a tool in and for assessment reform. Participants need no prior knowledge or experience with AI and, due to the hands-on nature of the workshop, will need to use their own laptops or tablets. It is envisaged that by the end of the workshop, participants will have gained a deeper understanding of the potential of AI in education, including a better understanding of its possibilities and challenges associated with its implementation.

# Moving to fully inclusive e-assessment

9:00 - 16:10

## Moving to fully inclusive e-assessment

H. Claydon[1], C. Jongkamp[2], T. Rousoulioti[3], I. Papakammenou[4]

[1]Freelance Assessment Consultant, United Kingdom
[2]Cito, Netherlands
[3]Aristotle University of Thessaloniki, Greece
[4]Centre of Foreign Languages Irini Papakammenou, Greece

It is often the case that diversity and inclusion are afterthoughts when an organisation is evolving its e-assessment offering. This workshop will provide an engaging opportunity for collaboration with peers, to consider the perspectives of a range of stakeholders. Thought-provoking discussions will equip participants with areas to take away and integrate in their future work practices.

The premise for the workshop is that participants work for a hypothetical assessment organisation that wants to update an onscreen assessment to make greater use of onscreen interactivity. The assessment has a diverse audience, including those with various access requirements.

Small group discussion will be used to explore the premise from multiple perspectives, drawing on experiences from across the group. The presenters will provide an informative introduction to various key areas, through short presentations, and participants will investigate some of these areas in more detail for themselves during the workshop and share findings with the group.

This workshop is led by members of the AEA-Europe eAssessment and Inclusive Assessment SIGs. We aim to provide participants with a basic understanding of inclusion considerations to inform everyday practice and/or opportunity to extend existing understanding.

No prior experience of e-assessment or inclusive assessment is needed.

# Responding As Assessment Professionals To Calls For Reform

9:00 - 16:10

## Responding As Assessment Professionals To Calls For Reform

A. Watts[1], E. Andressen[2]

[1]University of Cambridge, UK, United Kingdom
[2]Andressen Byram Ltd, United Kingdom

Education systems prepare learners for assessment and also assess them. But existing assessment practice in general, and academic education particularly, appears to define knowledge in a narrow way. Even allowing for the filtering purpose of compulsory examinations towards the end of schooling, some summative assessments put too much emphasis on knowledge recall.

This pre-conference workshop offers participants opportunities to reflect on their current assessment practice and how this is evolving. The workshop will explore three key themes: 1) What society expects of assessment professionals? 2) How to balance assessment of learning and assessment for learning? 3) Assessment for learning beyond schooling (e.g. in lifelong learning and for technical and vocational skills).

The workshop will invite discussion of, and learning from, existing practice. It will challenge participants to identify assessment issues and potential resolutions in a rapidly changing education context. Now is a time in which what it means to be employable is evolving and in which learning is increasingly experienced through technology and at a distance, even when education is delivered in traditional contexts. Attendees to the workshop will be able to explore how they might address the changing needs of students, of the workplace and of society more widely.

# Re)design YOUR assessment! Designing assessment tasks with evidence-centered design

9:00 - 16:10

## (Re)design YOUR assessment! Designing assessment tasks with evidence-centered design.

S. de Klerk[1], M. Waltman[1]

[1]Cito, Netherlands

Evidence-Centered Design (ECD) is a scientifically proven framework for streamlining the process of (re)designing your assessment tasks, and making your assessment practices more evidence-based (Mislevy, Almond, & Lukas, 2003). In this semi-hackathon workshop, participants from a multidisciplinary background take on the challenge to (re)design their own assessment practices, and make those more evidence-based, by working from an ECD perspective. Assessment redesign is increasingly important and necessary, as societal and technological changes require students, employees or citizens to demonstrate different universal KSA's (knowledge, skills and abilities) to be successful in school, a job, or society. On the one hand, we, as an assessment community, need assessment practices that match these changes in society. On the other hand, there is a need for assessments that fit the specific context of a country, educational institution, or company, and reforms that take place there. In both cases, there is a profound need to align what you want to evaluate, the task characteristics that allow insights into the what, and a measurement model for interpreting task performance and providing feedback and feedforward. In this workshop, we are going to work in teams on the challenge to build a blueprint for tomorrow's assessment, based on ECD.

## Keynote Speech

9:15 - 10:15

# MATSEC Examinations: An eventful journey

F. Ventura[1]

[1]University of Malta, Malta

In 1988, the newly-elected government of Malta decided to end the reliance of the education system on the GCE examinations offered by UK examination boards, namely, the Ordinary level examinations for certifying 16-year-old students at the end of compulsory education and the Advanced level examinations which were used for admission to the University of Malta. It set up a board to consider the full implications of this decision and the actions that needed to be taken to put it into effect. The board consisted of members from the University of Malta, the government Department of Education and the schools working as a partnership under the chairmanship of the Rector of the University. In 1991, this board was officially established as the Matriculation and Secondary Education Certificate Examinations (MATSEC) Board under the authority of the University Council through the Senate. The remit of the board was (a) to set up a system of examinations that would certify at least 80% of the cohort of students at the end of compulsory education, and (b) to replace the GCE Advanced levels by a system modelled on the International Baccalaureate Diploma system of examinations for admission to the university.

Naturally, these tasks presented formidable challenges regarding the required standards, equity, integrity and public confidence. Education officials openly doubted whether the university can have a structure to run these systems professionally. Their fears were based on the lack of personnel with expertise in educational assessment and the ‘small island state mentality’ where everybody knows everybody else giving rise to concerns about security, corruption, and nepotism. These concerns were also fuelled by the suspicion that the university would control the secondary school curriculum as the UK examination boards had effectively done for many years. Notwithstanding these doubts, also in 1991, the university set up a MATSEC Support Unit with an academic and an administrative division from members of staff to take responsibility for the day-to-day running of the examinations under the direction of the university Registrar.

After lengthy discussions internally, with the Ministry of Education and the Malta Union of Teachers, the Secondary Education Certificate (SEC) system of examinations for 16+ was launched in 1992 and the first new format examinations in 30 subjects were held in May 1994. These offered tiered papers, an extension of the grading scale to cover a wide range of abilities, and the introduction of oral examinations in languages and coursework in several subjects. The reaction by the schools, especially the private schools, was cautious as only 50.5% of the 16-year-old cohort sat for these examinations in the first year. This percentage rose gradually to 81.3% in 2005 and continued to exceed the projected 80% level since then. Indeed, this percentage exceeded the 90% level after the inclusion of vocational subjects at SEC level in 2014 with parity of esteem with the ‘traditional’ SEC subjects. This innovation took place following discussions between the University, the Directorate of Education of the Ministry of Education and the Malta College of Science and Technology (MCAST). A technical working group with members from the MATSEC Support Unit and MCAST discussed the format of the new syllabi and devised an interesting assessment scheme to ensure the acceptance of parity of esteem of the vocational subjects with the ‘traditional’ subjects.

Meanwhile, work on the replacement of the GCE Advanced levels by IB-type examination led to the launch of the Matriculation Certificate (MC) system of examinations in 30 subjects at Advanced level and 30 subjects at Intermediate level. Candidates for the MC needed to sit for two Advanced levels, three Intermediate levels and Systems of Knowledge also set at Intermediate level in the same session of examinations. The choice of

subjects had to include a language, a science, and a humanities or a business subject. This innovation was launched in June 1994 but the first MC examinations took place in 1997 since the schools needed more time to implement the necessary changes.

From a different perspective, the initiative of creating a local system of examinations at this level can be interpreted as an act of decolonisation. In effect, however, the colonial mentality, where the foreign product is always considered better than the local one, does not disappear by a simple declaration of independence. Several instances of the continued reliance on English sources ranging from the students' use of English textbooks meant for GCE subjects at Ordinary and Advanced levels in many SEC and MC subjects to references to the GCSE and GCE Mandatory Code of Practice and other documents for producing a local code of practice in the conduct of examinations and assessments.

The keynote address will elaborate on how the challenges were met in the first years of the operation of the MATSEC Examinations Board; the continued scrutiny by the schools, teachers, parents and the media; the introduction of vocational subjects; and the updates suggested in the evaluation reports published in 1998, 2005, 2010 and 2017

## Keynote Speech

10:45 - 11:30

# Investigating high achievement in mathematics and science in Ireland: An in-depth analysis of national and international assessment data

V. Pitsia[1]

[1]Educational Research Centre, Ireland

In Ireland, while, on average, students have performed well on national and international assessments of mathematics and science, the low proportions of high achievers in these subjects are noteworthy. Given these patterns and the multifaceted benefits in individual and societal terms that expertise in mathematics and science has been associated with, policymakers in Ireland have begun placing an increasing emphasis on high achievement in these subjects. This emphasis has coincided with ongoing efforts during the last decade to raise interest and improve academic performance within the realm of science, technology, engineering, and mathematics (STEM) education.
Despite this policy attention, research on high achievement in mathematics and science nationally, but also internationally, has been particularly scarce. In an attempt to provide research evidence that could add further impetus to the ongoing efforts, this study conducted an in-depth investigation of high achievement across education levels, student cohorts, and subjects using data from the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), the Irish National Assessments of Mathematics and English Reading (NAMER), and the Irish State Examinations (Junior and Leaving Certificates). The study aimed to (i) examine the magnitude and consistency of the issues related to high achievement, (ii) build profiles of high-achieving students, and (iii) evaluate the contribution of various contextual characteristics stemming from students, their families, teachers, and schools in the prediction of high achievement in mathematics and science in a multivariate and multilevel context.

The findings indicated that Ireland's percentages of high achievers and scores among students at the highest national percentiles of performance in mathematics and science have been significantly lower compared to countries with similar average performance. These issues, which were consistent across years and assessments, were more apparent for mathematics than science and at post-primary than at primary level, while similar patterns were not detected for reading. It was also found that variables related to students' self-beliefs, dispositions, engagement, learning approaches, and socioeconomic background were consistently associated with high achievement in mathematics and science. Overall, the significant role of students' homes and families in predicting students' chances of being high achievers in the two subjects was highlighted. In turn, this indicated that further efforts to enhance collaboration between teachers, schools, and parents may be warranted if progress in the area of high achievement in mathematics and science is to be made. The implications of these findings for policy and practice within the Irish context, the limitations of the study, and recommendations for future research will be discussed.

# Psychometrics and Test Development I

13:45 - 14:15

## How does DIF items affect the equating transformation?

M. Wiberg[1], I. Laukaityte[2]

[1]Umeå University, Sweden
[2]Umeå university, Sweden

Test score equating is used to make scores from different test forms comparable. It is common to use the nonequivalent group with anchor test design when equating test scores. Although we aim to avoid using differential item functioning (DIF) items we sometimes have them in standardized tests including the used anchor test. DIF occurs when groups with the same latent ability but from different groups have an unequal probability of answering a given item. The overall aim was to compare how different equating methods perform when we have DIF in either the regular test or in the anchor test. We used test forms from the Swedish Scholastic Aptitude Test (SweSAT), which is a multiple-choice binary scored test used for college admissions. We also used simulations based on the SweSAT to examine different conditions such as presence or absence of DIF and the effect size of DIF together with different features of the regular and anchor tests. Preliminary results show that the equated values vary depending on the amount of DIF especially when the anchor items have DIF. Practical implications and recommendations for how to handle DIF and lower its consequences when it appears in standardized achievement tests are given.

# Can examination papers always be error-free? An exploratory investigation into the conditions that can give rise to errors in assessment instruments

F. Constantinou[1]

[1]Cambridge University Press and Assessment, United Kingdom

Successful reforms presuppose adequate understanding of the underlying problem(s). In the area of high-stakes assessment, one problem requiring attention is the occurrence of errors in examination papers (e.g., multiple-choice questions containing more than one correct answer). Errors can undermine students' performance in the examination, while also eroding public confidence in the examination system. Therefore, it is important that they are prevented. In England, although various measures are taken to this end, errors continue to occur occasionally. This suggests that the current process of developing examination papers may need to be revisited. However, before any changes to the process are attempted, the conditions giving rise to errors need to be sufficiently understood. To investigate these conditions, this study analysed interview data from 36 assessment professionals involved in the examination paper construction process. The analysis exposed the complex and often unexpected ways in which the characteristics of the paper construction process and the characteristics of the people who participate in it interact with one another, creating opportunities for error. This talk will present some 'active' and 'latent' human failures which can lead to errors, and will conclude with some reflections on whether examination papers can always be error-free.

# An evaluation of targeting of items at assessment objectives in GCSE and A level qualifications in England

Q. He[1], Y. El Masri[2]

[1]Office of Qualifications and Examinations Regulation, United Kingdom
[2]Ofqual, United Kingdom

The General Certificate of Secondary Education (GCSE) and General Certificate of Education Advanced level (GCE A level) are academic qualifications awarded in specific subjects in England to students primarily aged 16 and 18 respectively. These qualifications generally contain multiple question papers developed based on specified content areas and assessment objectives (AOs) which represent the knowledge, skills and understanding that a student must demonstrate through responses to items in the papers. For each question paper, weights expressed as percentages of the maximum raw mark of the paper are allocated to different AOs to reflect the relative importance of different aspects of the assessed construct. Because the final observed marks are used to grade students, valid interpretation of the awarded grades requires that the original weights of the AOs are broadly achieved in the final observed marks. This presentation shares findings from research investigating the extent to which the intended weights of AOs in question papers from a selection of GCSE and A level qualifications from 2017 to 2022 are achieved operationally at item, paper and the overall qualification levels, the stability of the achieved weights over time, and systematic differences in achieved weights between different types of AOs.

# Educational Policy and Assessment in the era of decolonising curriculum I

13:45 - 14:15

## The Morality of Assessment

I. Nisbet[1], S. Shaw[2, 3]

[1]Faculty of Education, Cambridge University, United Kingdom
[2]Faculty of Education, University of Cambridge, United Kingdom
[3]Institute of Education, University College London, United Kingdom

This presentation will identify and analyse moral questions about assessment in educational contexts. Beginning with a brief explanation of the domain of “morality”, the presentation will focus on three questions: What is the moral justification for educational assessment? What moral challenges are made to educational assessment and are they persuasive? and, What moral principles should govern practice in educational assessment? The authors argue for a moral base for educational assessment and moral principles governing its implementation. The presentation will propose justifications for educational assessment as promoting or supporting private and public goods, but contingent on whether the benefits identified are achieved. Echoing the medical principle “First, do no harm”, the authors contend that that there is a moral imperative to minimise harm in when/how assessment is done and how its outcomes are used. They also consider how the rights of learners are affected by assessment. The presentation concludes that the most fundamental moral questions underpinning educational assessment in the mid-21st century are: “What evidence (from the ever-increasing mass available) is it right to use to make judgements that matter about learners?”; and “What processes are morally justified to obtain that evidence, make the judgements and communicate the outcomes?”

# Professional Testing Guidelines as Tools for Improving Educational Assessment: The Role of the International Test Commission

S. Sireci[1], T.T. Nguyen[1]

[1]University of Massachusetts Amherst, USA

Educational assessments are popular components of educational reform efforts, in part because they are seen as providing objective information to inform educational policy and instruction. However, measuring the knowledge, skills, and capabilities of children is difficult, in large part due to the diversity of personal characteristics and educational experiences inherent in schools within and across nations. Several national and international organizations have developed guidelines and standards to assist test developers in creating, administering, scoring, and reporting test results that are suitable for all children and provide valid and reliable results to fulfill the purposes for which the tests are developed. In this presentation, several ITC guidelines will be reviewed, the most recent of which is Guidelines for Technology Based Assessments, which were co-produced between the ITC and the Association of Test Publishers (2022). We will provide an overview of the ITC and its role in producing guidelines for promoting fair, appropriate, and efficient assessments across the globe, and compare ITC guidelines with other national and international guidelines and standards for promoting quality assessment. The importance of professional guidelines for testing, and suggestions for involving the Association for Educational Assessment-Europe in future guideline development efforts, will be discussed.

# The journeys of large-scale assessment systems from an international perspective: towards a formative, low-stakes, democratic, contextualised, and holistic approach

M.T. Florez Petour[1]

[1]Pedagogical Studies Department, University of Chile, Chile

The study on which this paper draws was aimed at nurturing discussions about the future of Chile's assessment and support system after COVID-19, from an international perspective. It comprised an international review of large-scale assessment and support systems whose design responded to a series of criteria that detached from the logic of performance-oriented testing and accountability systems, of which Chile is a paradigmatic case. The study considered two phases of research. An initial stage involved constructing a corpus of 28 cases in different geographical zones based on six criteria: i. formative purpose; ii. low-stakes; iii. underlying pedagogy; iv. participation; v. holistic approach to education; vi. consideration of context and diversity. A general characterisation of the cases was developed. A second phase of the study involved the selection of a sub-sample of 10 cases for in-depth review on the basis of research and policy sources and semi-structured interviews with key actors. This review provided a more nuanced understanding of the cases, including obstacles and facilitating aspects, along with context-related factors and tensions. Conclusions offer lessons from each case for the transformation of assessment and support systems in Chile and internationally, particularly relevant in the current scenario of uncertainty and change.

## Summative Assessment I

13:45 - 14:15

### High stakes assessment that supports mathematical problem solving: a journey of realistic aspiration or of chimera?

J. Golding[1], B. Redmond[2], G. Grima[3]

[1]University College London Institute of Education, United Kingdom
[2]Pearson, United Kingdom
[3]Pearson UK, United Kingdom

From the first national curriculum in England in 1988, successive mathematics curricula for ages 5-16 have attempted to promote robust mathematical problem-solving, reasoning and communication for all. This aspiration, although enjoying broad support, has remained largely unrealised. The curriculum introduced from September 2014 again included a renewed focus on these key mathematical processes - this time within an increasingly high-stakes assessment system. National assessments in England are developed within a marketized system; they include GCSE Mathematics, taken by nearly all students at age 16, and high stakes for them as individuals, for their teachers, and for their schools.

We draw on three secondary school curriculum enactment studies of this curriculum and its assessment, two predating the covid pandemic and one ‘New Normal’ study probing emerging practices and learning post-pandemic. These all harness classroom observations and related teacher and student voice. We uncover a story of repeated attempts to support development of these processes via resources and assessment reforms initially well-aligned with intentions. We analyse the interdependent challenges of doing so – for assessment organisations, producers of curriculum resources, students, schools, teachers, and policymakers. We argue that systemic changes are needed if mathematically laudable aspirations are to be realised.

# Investigating how access to statistical evidence and teacher feedback influences examiner judgement in grade awarding.

L. Badham[1]

[1]International Baccalaureate, United Kingdom

Different sources of assessment evidence are reviewed during International Baccalaureate (IB) grade awarding to convert marks into grades and ensure the fairest results possible for students. Qualitative and quantitative evidence are analysed to determine grade boundaries, with statistical evidence weighed against examiner judgement and feedback on examinations. A trial was conducted to explore how examiners' grading of examination scripts was influenced by having access to statistical evidence and teacher feedback on examinations. Grade awarding processes were simulated in five subjects, with examiners accessing all assessment evidence in one condition and only script evidence in the other. Grade boundary recommendations were compared in each subject, and focus groups were carried out with examiners and assessment staff to gather feedback on the advantages and disadvantages of the different approaches. Feedback from subsequent focus groups indicated that script evidence was central to examiners' roles in awarding. However, whilst awarding participants reported that item-level data was essential for identifying and prioritising questions for review during grading; the disadvantages appeared to outweigh the advantages for examiners accessing other forms of evidence during the awarding process. These findings suggest it may be more beneficial to share such evidence with examining teams outside the awarding process.

## International Assessments

13:45 - 14:15

# Applying Differential Item Functioning analysis to evaluate the comparability of language versions of PIRLS in South Africa

H.L. Kayton[1]

[1]University of Oxford, United Kingdom

PIRLS is currently the only large-scale assessment of reading conducted in all eleven official languages in primary schools in South Africa. In 2016 students performed poorly, and when comparing PIRLS achievement across language groups, the results vary considerably. Almost half (43%) of students who wrote the test in English demonstrated an ability to read at a basic comprehension level, while only 7% of students who wrote the test in Sepedi reached the same level. This study investigated the extent to which the English and Sepedi language versions of PIRLS Literacy 2016 could be considered comparable at an item level. An Item Response Theory based likelihood-ratio test approach to Differential Item Functioning (DIF) detection was used to identify items that function differently across groups. Item functioning was evaluated in terms of both difficulty level and discrimination ability. In terms of difficulty, 24% of items functioned differently, and in terms of discrimination 11% of items functioned differently. Added to this, there were several issues found with the alignment between item difficulty and student ability for the Sepedi group. Overall, the findings suggest that there is evidence to suggest that PIRLS Literacy 2016 does not function comparably for both groups.

# Pre-smoothing and other approaches to linking scores between mixed tier assessments.

B. Ashworth[1], L. Liu[1], B. Donahue[1], P. Pirc Zagar[1], S. Nastuta[1]

[1]Pearson, United Kingdom

Assessment scores for subjects with mixed tiers, which vary in difficulty, need to be adjusted or linked to ensure that candidates on a harder paper receive a better outcome than a similarly scored candidate on an easier paper.

Score linking is carried out by the chained equipercentile method (CEP) which also reflects the overall common item performance for each tier.

Method reliability maybe compromised when overall mark distributions are unsmooth, particularly for subjects with lower candidate numbers, and a pre-smoothing method is proposed using a loglinear model of suitable degree. A degree is selected, via goodness-of-fit analyses to ensure sufficient parameters are available for obtaining the best possible fit and to be used consistently for all scenarios.

Despite goodness-of-fit results showing improvement for higher degrees, there is less impact from unsmoothed distributions and less-populated cohorts, plus literature reviews suggest degree 3 to 6 models are suitable for most cases. Pre-smoothing provides a better reflection on the overall distribution, from smaller represented subjects, and increased confidence on the validity of linked grade boundaries for both tiers. Other literature reviews suggest ‘Nominal weighting’ are more accurate for significantly smaller entry subjects, and Item Response Theory (IRT) methods can improve accuracy of tier-linking.

# Implementing the ISAWG Method of Standard Setting and Maintenance in the International Baccalaureate

C. Hope[1], B. Smith[2]

[1]AlphaPlus Consultancy Ltd, United Kingdom
[2]AlphaPlus, United Kingdom

The International Baccalaureate (IB) is a major international non-profit foundation which offers a suite of educational programmes to students aged between 3 and 19. As a result, one of the myriad of tasks for the IB in their programmes' running is setting and maintaining the standard of these assessments, which they considered reforming to ensure optimal fairness and comparability from year to year.

Statistical evidence is one strand of evidence that IB uses in standard setting and maintenance, and IB commissioned AlphaPlus to investigate statistical approaches to setting grade boundaries. Whilst a range of approaches were investigated, in this presentation we will focus on the ISAWG (Instant Summary of Achievement Without Grades) method due to its utilisation of a broad range of subject data to acquire an overall measure of candidates' general academic ability.

ISAWG appears to be a viable form of statistical evidence for standard setting in an IB context, particularly for examinations where the number of candidates is in rapid flux. As such, this research was an important part of IB's assessment reform journey and has paved the way for discussions about how such an approach could be implemented in practice.

# Assessment that is reactive to unforeseen circumstances (e.g. Covid 19) I

13:45 - 14:15

## Global and Intercultural Skills Program: Intercultural Perceptions Student Index

T. Milford[1], V. Glickman[1], J. Anderson[1]

[1]University of Victoria, Canada

In British Columbia (BC), the Ministry of Education has recently recognized the importance of intercultural skills and global competencies. In an attempt to address these skills and competencies, the Ministry is piloting a Global and Intercultural Skills Program (GISP) for students. Although the program involves collaboration across a number of school districts, a clear, valid, and reliable assessment instrument associated with this program is lacking. This presentation details the development of a demonstration assessment tool that can be used with the GISP. Specifically, the Intercultural Perceptions Student Index (ICPSI) is intended to provide information about student perceptions of school-related experiences and learnings in the global and intercultural domain. The items composing the ICPSI were purposefully selected based upon a review of GISP documentation from the Student Learning Survey which has been administered to all BC students in grades 4, 7, 10 and 12. The selection of items in the ICPSI was based on a relationship to global and intercultural learnings, and to exposure and awareness of Indigenous cultures. In this presentation, we introduce the questions that make up the ICPSI, describe some of the psychometric properties, and make recommendations for how the ICPSI might best be used in BC.

# Evolving understandings: A longitudinal analysis of teacher candidates' approaches to assessment

C. Schneider[1], C. DeLuca[2], L. Müller[1], A. Coombs[3]

[1]University of Trier, Germany
[2]Queen's University, Canada
[3]Memorial University, Canada

Competence in assessment is a priority outcome of teacher education (TE) programmes, and numerous reform efforts have tackled embedding assessment learning in TE. Learning about assessment is a complex, social endeavour, involving theoretical input, practicums, and reflection of learning. Previous research has found that learning about assessment requires sufficient time. Our paper works on the research question of which specific facets of teacher candidates' assessment competence evolve throughout TE.

In a longitudinal survey design with 394 student teachers inside a TE programme, the Approaches to Classroom Assessment Inventory was administered twice, 2 ½ years apart. The ACAI seeks to determine teacher candidates' approaches related to four themes of assessment: purposes, processes, approaches to fairness, and theory.

Latent change modelling on a high aggregation level reveals that particularly, 'contemporary' approaches to assessment require time to develop during TE. Manifest analysis on single aspects shows that candidates' focus shifts towards formative assessment and develops away from 'standard' approaches in fairness (equal criteria for all) towards more equitable/differentiated views. Our data support that student teachers' learning about assessment is complex and occurs at slow pace. Hence, future "reform journeys" may be well-advised to refrain from 'speeded' programmes when it comes to assessment learning.

# E-Assessment I

13:45 - 14:15

## What can process data can tell us about students' persistence? Evidence from the e-TIMSS 2019 assessment

E. Papanastasiou[1], E. Konstantinidou[1]

[1]University of Nicosia, Cyprus

The digital transition in large-scale assessments generated a plethora of log-data useful for testing the validity of test scores use and interpretation. The study aims to describe and evaluate two indicators of examinee test-taking persistence (i.e., successful persistence and unsuccessful persistence) and examine their relationship with student performance in mathematics and science. Large-scale, representative data from e-TIMSS (Trends in International Mathematics and Science Study) 2019 and from 4th graders in the USA who participated in both mathematics and science assessments were used. The predictive validity of those indicators that investigate students' persistence with respect to achievement estimates was examined. Based on the preliminary results of the study, the unsuccessful persistence indicator was more highly correlated with achievement, indicating that students with lower levels of achievement ended up spending more time than average on items that were eventually answered incorrectly; whereas higher achieving students had lower amounts of unsuccessful persistence in incorrect answers. Our study departs from examining only whether a student answered correctly or incorrectly and sheds light on the process of answering an item, including the degree of persistence an examinee demonstrates.

# A new quantile regression approach to age-standardisation for on-demand assessments

M. Turner[1], B. Smith[1]

[1]AlphaPlus, United Kingdom

Many current high stakes summative assessments rely upon a mandated sitting window, which is described as ‘fair’ because everyone has equal preparation time. However, this is arguably not fair, as anyone with a short-term impairment (e.g. flu) is hugely impacted; clearly on-demand e-assessment can be fairer. The question then becomes how to validly compare outcomes given different sitting times.
Norm-referenced standardised scores are one solution and provide information about where learners sit relative to others in the cohort. However, one issue with standardised scores is the age effect; at a given point in time, younger learners tend to perform less well than older learners. This is especially pronounced for the youngest year groups, where up to twelve months of cognitive development can have a very real impact.
This led to the development of age-standardised scores, which correct for the effect of age. Current age-standardised methodology relies on segmenting the data into “month of birth” groups and fitting a model to each sub-group. However, this does not take into account time of sitting within an academic year. This presentation details a novel quantile regression-based methodology for carrying out ‘age and learning time’ standardisations which has numerous benefits over other approaches.

# The reform journey of an on-screen national assessment

A. Boyle[1]

[1]AlphaPlus Consultancy Ltd., United Kingdom

Wales runs a suite of on-screen assessments, delivered using an adaptive model. The principal purposes of these assessments are formative, rather than summative or accountability.

This system has now been running for several years. The model is appropriate – given very wide ability ranges in any school year. Further, feedback on the assessments to date has been positive.

However, as we progress, we sense that the adaptive assessments are more positive for learners at the lower end of ability distributions. Questions tailored at their (lower) ability gives them a more positive assessment experience.
We have some feedback that higher ability learners find these new assessments more challenging. This is particularly so for younger learners. Further, there is a strong sense that ‘it isn’t fair to test content which children haven’t been taught’.

In this presentation, we will discuss the amendments to assessment designs, which we have agreed with a teacher group. To some extent, this means abrogating the ‘purist’ adaptive model. But this discussion with teachers has also facilitated an understanding of fundamental issues in learning and assessment, how professionals can arrive at differing interpretations, and yet how differing views can be reconciled.

# Comparative Judgement

13:45 - 14:15

## Multiple marking using the Levels-only method in A level English Literature

E. De Groot[1], J. Ireland[1]

[1]Cambridge University Press and Assessment, United Kingdom

Essays capture constructs that other assessment items cannot. Currently to mark GCSEs and A-levels, examiners allocate marks, nested within levels of performance, for different assessment objectives (AO). This process is time intensive, typically one examiner marks each script. Thus, it potentially risks rendering the marking less reliable as dependent on an individual examiner's preferences. Whilst this can be mitigated with standardisation, seeding, and statistical scaling, more subjective subjects and essays still risk having lower marking reliabilities.

To address this a less time-consuming "Levels-only (LO)" marking method was developed where the essay is marked using only the levels for each AO in the existing scheme. Marks within levels are not allocated and no annotations are given. When tested on a short essay from GCSE English Language, triple LO marking was encouraging, yielding high reliability and predictive values whilst using similar examiner time as traditional marking.

This new research investigated LO marking with longer A-level English Literature essays. The marks were compared with the original live marking and similarly to the previous study finds double marking can be done with high reliability and predictive values faster than traditional marking. Finally, we discuss feedback from the examiners about their experience of the method.

# Experiences from reforming the math exams in Norway

B. Vinje[1], O. Tokle[2]

[1]National Centre for Mathematics Education, NTNU, Norway
[2]NTNU, Norway

In the period 2020 – 2022 the Norwegian Centre for Mathematics Education assisted The Norwegian Directorate for Education and Training in developing new math exams for different math subjects in lower and upper secondary school. We got a new curriculum in Norway from 2020, and with the new exams the directorate wanted to change both the system for completing the exams and the math content in the exams.

In this presentation we will look at and discuss all the obstacles and challenges we met while developing new exams, mainly connected to the work with a new exam for mathematics in 10th grade. How did we handle a continuous change in guidelines from the directorate? Which changes did we do in terms of content to reflect the new curriculum in the best possible way? And in the end, how should we assess the student's mathematical competence? Developing the new exams was demanding, but we learned a lot, and the developing process reflects many of the considerations decision makers and test developers need to consider when reforming assessment.

# Linking two scales using comparative judgement

A. Béguin[1], E. Crompvoets[2], M. Van Onna[3]

[1]IBO, Netherlands
[2]Tilburg University, Netherlands
[3]Cito, Netherlands

This paper proposes a procedure to use comparative judgement to link the results of two test that are administered to different populations. In this procedure judges are asked to compare the difficulty of items from the two tests. And more specifically which item has a higher probability of being answered correctly taking into account the perspective of a student with a certain proficiency level. It is explained why it is important to take into account the proficiency level of the student. We compare data collected using different instructions (1) without mentioning proficiency, 2) only focusing on the average student 3) with focus on low, average and high proficiency).
We give specifics of the estimation procedures of mean-sigma linking based on the data collected with each of the three instructions. We evaluate the efficiency of each of these procedures based on a simulation study and we provide data based on a small-scale study using the above instructions.

# Technical, Vocational and Applied Assessments I

13:45 - 14:15

## Towards understanding the quality and value of outcomes-based qualifications: academic criticisms versus lived practices of awarding organisations

M. Curcin[1], A. Brylka[1], L. Clarke[1], P. Newton[1]

[1]Ofqual, United Kingdom

This talk focuses on a long-standing approach to qualification design, underpinning many vocational and technical qualifications in England, which we call ‘CASLO’. It defines qualification content and standards in terms of detailed learning outcomes and assessment criteria and requires learners to achieve all specified learning outcomes to pass.

The existing literature is largely critical of this approach, with some suggestion that it may be inherently sub-optimal. But the literature is also dated, fragmented and polarised. From it we identified a taxonomy of potential assessment, teaching, learning and delivery problems and invited awarding organisations (AOs) delivering CASLO qualifications to reflect on them based on their exemplar qualifications during semi-structured interviews. We asked whether they recognised each problem, what mitigations they implemented if so, or why any problems might not apply in their context. Our sample included 15 qualifications across 14 AOs, several levels, subject areas, types and uses.

We present aspects of our analysis of the AO responses and draw conclusions about the implications this has regarding the force of the original criticisms given AO design decisions, uses of results, target cohorts, situational contexts and the near-universal challenge from the AOs to the suggestion that the approach might be inherently sub-optimal.

# Assessing ‘competence’ in education reform projects – what lessons can we learn from technical and vocational education?

R. Conway[1, 2], M. Dean[3]

[1]Freelance Assessment Consultant, United Kingdom
[2]Consultant, Cambridge Assessment International Education, United Kingdom
[3]Cambridge Assessment International Education, United Kingdom

‘Competence’ has become a common feature of learning and assessment frameworks and curricula. In our organisation, we are seeing significant demand from clients engaged in education reform projects that focus on competence-based curricula and assessment. To support the implementation of this work, we are developing principles to underpin quality and consistency in competence assessment across a range of different national contexts and requirements.

Despite the demand for work on competence-based curricula and assessment, the concept of ‘competence’ is not self-evident and various definitions have been proposed in different education policy and practice contexts.

While competence assessment is emerging in general education, it is well-established in technical and vocational education. Technical and vocational education therefore provides a useful prism through which to explore competence assessment more broadly, particularly in areas such as assessment methods and approaches to delivery. Given the recent reforms in England, we use this as our primary case study.

In this presentation, we draw upon research, policy, and practice to interrogate what is meant by ‘competence’ and identify ‘lessons learned’ to inform principles for developing competence assessment in education reform projects.

# Exploring the quality and value of vocational qualifications in England: reflections from students, teachers, employers and higher education recruiters

L. Clarke[1], P. Newton[1], M. Curcin[1], A. Brylka[1]

[1]Ofqual, United Kingdom

Vocational and technical qualifications in England have been subject to a number of reforms over the past three decades. Despite changes across the landscape, there is a core group of qualifications which have continued to exist. These qualifications adopt a specific approach to qualification design which we call the 'CASLO' approach. They can be recognised by their detailed learning outcomes, explicit assessment criteria and mastery approach to assessment.

As part of a larger programme of work, Ofqual has interviewed several Awarding Organisations (AOs) to better understand design principles associated with CASLO qualifications. These explorations unearthed a range of decisions made at the design phase, however the impact of these decisions for users are unclear. Therefore, we conducted focus-group-interviews with 57 teachers, students, employers and higher education recruiters to explore a range of perspectives regarding the quality and value of CASLO qualifications.

Findings revealed that stakeholders benefitted from many of the design decisions made by AOs. However, trends in the data also suggested that there are less desirable aspects of CASLO qualifications including burdensome assessment processes and differences in centre and industry standards. Furthermore, there are differences relating to implementation which influence and shape views on the robustness of these qualifications.

# Formative Assessment I

15:45 - 16:15

## Self-assessment in English as a foreign language Students' written self-assessments and students' and teacher's reflections

A. Gillespie[1]

[1]Oslo Metropolitan University, Norway

Self-assessment is an important feature of assessment for learning. To investigate how self-assessment is carried out in the subject of English as a Second Language we conducted a case study where we collected all self-assessments written by a group of Norwegian 8th grade students in the subject of English Language over the course of one year. (N=19). Moreover, we conducted focus-group interviews with the students and one interview with the teacher. The findings suggest that the better part of the written self-assessments deals with artificial features of the student's performance and to a larger extent issues related to classroom conduct and effort. In the interviews the students claimed to be unfamiliar with the rationale for doing self-assessment. The students reflected upon the self-assessment practice, and expressed that they perceived engaging in self-assessment as "useless" and "pointless". We suggest that students need support in developing self-assessment skills and call for more research concerning how to best develop these skills, along with more research that involves the students' perspectives.

# Journeys of self- and peer-assessment in a reformed mathematics curriculum: primary school children's accounts of the roles of explanation, reflection and challenge.

G. Grima[1], J. Golding[2], B. Redmond[3]

[1]Pearson UK, United Kingdom
[2]University College London IOE, United Kingdom
[3]Pearson, United Kingdom

The 2014 primary curriculum in England has an ambitious focus on genuine mathematical problem-solving, reasoning and communication. Such aspirations have previously proved intractable. We draw on primary school children's voice from a longitudinal (2019-2022) classroom-close study of the use and impact of one set of teacher-educative curriculum materials, 'Power Maths', to show how reformed pedagogical devices and deliberate, semi-structured and probing class discussion, if used consistently, can support children's peer- and self-assessment for curriculum intentions and the development of productive mathematics dispositions.

Visits to year 2, 4 and 6 classes (age 6/7, 8/9 and 10/11) gave access to children's views on their 'new normal' post-pandemic practices/views?. We focus on responses to two Power Maths devices, 'Reflect' and 'Challenge', and also sample children's analysis of the reformed role of in-class mathematical explanations. We show children were keen and able to articulate the learning potential, as well as inherent demand, of such approaches – and that almost all embraced that/them?. In two schools where teachers had invested heavily in knowing the linked teacher-educative support materials, the quality of children's mathematical communication was exceptional. We suggest that well-structured curriculum resources supporting active learner self- and peer-assessment can promote achievement of aspirational curriculum intentions.

# Changing assessment cultures and practices through e-learning

V. Meland[1], E.W. Hartberg[1]

[1]Inland Norway University, Norway

«I have almost turned 180 degrees. From results and grades to focus on the students learning process. » - Teacher

In this paper we will focus on how teachers and school leaders develop and change their assessment cultures and practices through organizational based e-learning courses. The Norwegian government has recently changed the curriculum for schools, and the enactment for formative and summative assessment have also been adjusted to correlate with the new curriculum.

To support teachers and school leaders in implementing and acting upon these changes, the Norwegian Directorate for education and training initiated several organizational e-learning courses that was developed by Inland Norway University of Applied Sciences together with other professional educators representing different universities in Norway.

This qualitative study explores different aspects of changing the assessment cultures and practices at two schools through the e-learning initiative. Therefore, we have raised three research questions that we will address in this paper:

1. What are the main changes in the assessment enactment and the new curriculum?
2. How do teachers and school leaders learn and collaborate in the e-learning courses about assessment and the new curriculum?
3. What are the implications for the assessment practices and cultures at these schools?

# E-Assessment II

15:45 - 16:15

## Computer-based tests and machine marking: candidates' perceptions and beliefs about test taking experiences

M. Richardson[1], R. Clesham[2], S. Leaton-Gray[3]

[1]UCL Institute of Education, United Kingdom
[2]Pearson UK (corporate membership), United Kingdom
[3]UCL IOE, United Kingdom

When considering reform in assessment research, it is important to explore the role of recent developments particularly Artificial Intelligence (AI) and its use globally in education. The research presented here explored how candidates understand and interact with a language test that uses AI – with the aim of documenting and explaining their lived experiences; and determining their technical understanding and beliefs about AI-led tests.

Research about the use of AI in testing remains largely focused on the technical and this study builds on this by adding the personal experience of the candidate in an AI language testing setting. Data were collected with an online survey (n=486); and individual interviews (using Teams, n=21). Three overarching categories came out of the analysis: Test preparation, Test taking experiences, and Perceptions of AI.

The results reflect ways that candidates interact with the online nature of the tests and reveal assumptions, misconceptions and beliefs about the nature of AI in testing, and particularly how AI is used in these tests. The presentation shows examples of feedback from test takers and will consider how test developers might further acknowledge and interrogate the lived experience of candidates interacting with AI-led tests.

# Towards external assessments dematerialization – Are we ready? Portuguese school principals' concerns.

G. Cipriano[1], S. da Cruz Martins[1]

[1]CIES-Iscte, Portugal

In 2020, the confinements caused by the Covid-19 pandemic involved the displacement of external large-scale assessments in favour of assessments built by teachers. In a post pandemic era, the reintroduction of external assessment was again discussed in Portugal. In addition to external assessment reintroduction, the project for External Assessment Dematerialisation (DAVE) foresees national gauging tests and exams in digital format. Assessment reforms, such as DAVE, brings concerns to school communities about its implementation and reliability. To identify and analyse those concerns, we have conducted 32 semi-structured interviews with school principals from mainland Portugal. Results showed that, on the one hand, there are some school principals that state that schools are ready, and DAVE's implementation is inevitable in a near future. On the other hand, a major part of school' principals considers that schools are not ready and there is a lack of investment before its implementation. Apprehensions about schools' technological capacity, internet access, regular use of ICT for teaching, average teachers' age without a concerted and continuous training plan for the application of digital educational resources in teaching, create fears of social injustice and alarms about the fairness of DAVE, especially when external assessments have stakes associated for students.

# Response Model Validation in Digital Mathematics Assessments

F. Salles[1], B. Maddox[2], S. Keskpaik[1], P. Boon[3], M. Meangru[4]

[1]DEPP, Ministry of Education, France
[2]Digital Education Futures Initiative, Hughes Hall, University of Cambridge, United Kingdom
[3]Numworx, Utrecht University, Netherlands
[4]University of East Anglia, United Kingdom

Since 2017, technology-enhanced on-screen test items (TEI) have been included in large-scale assessments of mathematics in France. TEIs can improve the assessment of higher-order skills and capture response processes of respondents within the testing platform. The use of technology in the mathematics class is required in the national curriculum in France, and a rich assessment environment involving the use of digital tools improves the validity of the test in relation to the curriculum as well as to the mathematical concepts at stake. The DEPP, along with other institutions, has developed a methodological framework for the use of log data in large-scale assessments based on the interaction of didactical, technological, and analytical aspects. This paper presents a mixed-method approach that includes the development and application of statistical models to analyze log data, and complementary empirical observations involving eye tracking and video analysis of gesture, talk, and facial expressions. The aim is to refine the descriptions of interactions with test items and digital tools, inform on item design and improvements to user experience, and contribute to response models development and validation.

# Fairness & Social Justice I

15:45 - 16:15

## Fair and effective? Staff and student perspectives on 25% extra time in exams in England

L. Kennedy[1], S. Holmes[1]

[1]Ofqual, United Kingdom

SEND (special educational needs and disabled) students in England are legally entitled to reasonable adjustments to their exams, of which 25% extra time is the most frequently allocated adjustment in England (Ofqual, 2022). This figure has been steadily rising over several years, attracting concern from British assessment bodies. Two studies were undertaken by Ofqual to understand the perceptions and experiences of extra time. The first consisted of interviews (n = 15) with educational staff and secondary school students, the second included a survey (n = 387) and interviews (n = 15) with students who had recently taken high-stakes exams. Teachers reported that students who may benefit from extra time were generally identified by teachers, but the level of teacher experience impacted the skill to do so. Limited resources available to schools and colleges affected ability to gather evidence for and apply for adjustments, and if this causes delays, the student has less time to practise using extra time. Students generally thought extra time was a fair adjustment for those who needed it but fears regarding unequal access between schools were raised. There is an overall view that exams are rather time pressured, varying by subject.

# The future of accessibility in assessment: practitioners' views on current and future access arrangements in England

K. Finch[1], P. Surridge[1]

[1]AQA, United Kingdom

Ensuring all groups of learners have fair and equitable access to assessments is a legal requirement in England and a key consideration for awarding organisations. For some students, this involves access arrangements being put in place, such as modified exam papers or the provision of extra time to complete an exam.

In England, the practitioners in schools who facilitate these adjustments include special educational needs coordinators (SENCos) and exams officers. This paper qualitatively examines the views and experiences of a sample of these practitioners (n=17) with reference to both current access arrangements as well as how the future of accessibility looks for their learners.

Through reflexive thematic analysis of interview and focus group data, three main themes were generated: patterns in access arrangements; factors affecting access arrangements; and accessibility in on-screen assessment.

This paper will unpack subthemes in the data, which include the increase in requests for extra time, how social, emotional and mental health conditions impact exam accessibility, the growing role of parental influence, and the potential benefits and challenges of moving to on-screen exams.

## Assessment Cultures I

15:45 - 16:15

### Applying policy learning from two cultures to a third. What developing and evolving skills-based programmes can really learn from Germany and England about technical and vocational education and training (TVET) structures, practice and assessment: insight

E. Andressen[1], S. Shaw[2, 3]

[1]Andressen Byram Ltd, United Kingdom
[2]Faculty of Education, University of Cambridge, United Kingdom
[3]Institute of Education, University College London, United Kingdom

This presentation identifies key elements and characteristics of the England and German technical and vocational education and training (TVET) systems which have been explored in the context of their ability to support the 2022 policy reforms announced with the intention of reforming Chinese vocational provision. The presentation draws on the literature around policy transfer, and the origins of the English and German technical and vocational education and training systems. It explores the benefits and limitations of two well-established systems in terms of their ability to offer features and practice which can be adopted or adapted in a third national context. The research uses opportunistic data from business projects with China, to provide support in the areas of system reform, capacity building, stakeholder engagement, delivery and assessment models, and quality assurance. There are many potential models which can serve those wanting to increase skills acquisition but transferring wholesale from one system to another is unlikely to succeed without adaptation. Local and national cultures in both the home and source jurisdictions need to be understood, navigated and eventually evolved, to create a system uniquely fitted to meeting current Chinese ambitions and context.

# Enacting assessment reform in Cox’s Bazar Refugee Camp: a case study

G. Billings[1], S. Nelson[1]

[1]Cambridge University Press and Assessment, United Kingdom

In Cox’s Bazar, Bangladesh, a network of refugee camps operates for the displaced Rohingya from Myanmar. Just under a million people live in these camps, and around half are children. Most Rohingya here have no legal identity or citizenship and are entirely reliant on humanitarian assistance .
Cambridge University Press and Assessment were asked by UNICEF to create an assessment framework for children being educated in the camps, alongside a system for monitoring and evaluating its implementation and impact. This was planned against the background of a whole-scale transition from an emergency curriculum (LCFA) to the Myanmar Curriculum in learning centres.
The enactment was complex, influenced by Covid-19, stakeholder engagement, and the changing geo-political landscape. Further cultural challenges, such as the engagement of girls in education, and the significant language barriers between teachers also had to be considered.
This presentation will focus on the enactment journey, with details from our experience in the refugee camps with planning, implementing and evaluating the project, updated to reflect our experiences over the next seven months of the project.

# Assessing Learning Outcomes in Finnish Basic Education: Critiques and Challenges

J. Marjanen[1], M. Huhtanen[1]

[1]Finnish Education Evaluation Centre, Finland

National and international assessments have revealed a decline in Finnish pupils' competence since the early 2000s. The evaluations of learning outcomes in basic education are sample-based, with the main focus being on the overall competence of the entire age cohort, rather than the individual pupil. Additionally, the emphasis is on the conditions for learning as well as equality and equity in the Finnish basic education.
It has been criticized that the current assessment system cannot provide sufficient information for policymakers on how to reverse this negative trend.
Some have suggested that natural experiments and quasi-experimental designs (e.g., instrumental variable analysis, regression discontinuity designs, and propensity score matching), national standardized tests, and taking a wider macro-level view of the societal changes that affect schooling, may produce better knowledge of the causes behind changes in pupils' competence over time.
Even though they do not seem to provide easy answers, broadening the scope of educational assessment and the variety of methods used is necessary in the ever-changing world if policymakers are to be provided with more solid information on how to develop the school system in the future.

## National Tests & Examinations I

15:45 - 16:15

### Evaluating the impact of curriculum and assessment reform in secondary education on progression to mathematics post-16

C. Vidal Rodeiro[1], J. Williamson[1]

[1]Cambridge University Press & Assessment, United Kingdom

In England, GCSE (General Certificate of Secondary Education) qualifications offered to students aged 14-16 were recently reformed with the intention of enriching the curriculum and better preparing students for future education or employment. For mathematics specifically, the new GCSE aimed to be more demanding, provide greater challenge for the most able students, and support progression to post-16 mathematics. However, there have been concerns that the new GCSE could deter students from post-16 study (e.g., by reducing their confidence) and, to date, there has been little research on its impact on participation in and learning of mathematics post-16.

The current research approached the question of how the reform of GCSE mathematics affected progression to and performance in post-16 mathematics and maths-related subjects via quantitative analysis of entries and performance data pre- and post-reform.

This research has raised important issues for the mathematics education community and for policy makers by increasing the understanding of how recent reforms to upper secondary mathematics have affected students and contributing evidence on progression to post-16 study. In particular, it has shown that participation in mathematics and maths-related subjects post-16 generally increased following the reform. By contrast, compared to the pre-reform years, performance was generally worse.

# Can centralising marking in Sweden improve interrater reliability?

J. Anker-Hansen[1], D. Gustafsson[1], C. Johansson[1], N. Ekblom[1]

[1]Swedish National Agency for Education, Sweden

The Swedish school system has a tradition where teachers have marked the standardised exams of their own students and decide themselves how the results should influence the students' grades. Recent governmental reports have created a political sensitive situation concerning how reliable grading in Sweden really is.
The Swedish National Agency for Education was tasked to evaluate if centralising marking of standardised exams with trained markers could lead to an increased interrater reliability. Two testlets in Swedish and one in English with approximately 300 student responses were collected and distributed among 24 markers per testlet. A third of the markers were the students' own teachers, a third, teachers from other schools and a third, Agency trained expert markers. All markers received 10 common control responses in their quota that were used for computing consensus and consistency estimates.
Neither consensus nor consistency estimates differed between the three groups regardless of testlet. Hence, the results did not support the political suggestion to improve reliability of marking standardised exams by centralising marking. Possibly, some items in the Swedish standardised exams were too open ended to train markers successfully for the vast variety of responses.

# Implementation of the use of central tests in Flanders. Teachers' readiness to use formal performance data to improve student learning and the impact of school culture.

G. Molenberghs[1], J. Vanhoof[1], R. Van Gasse[1]

[1]Antwerp University, Belgium

Although the empirical evidence on the educational impact of a systematic use of formal performance data (from central tests, among others) is quite strong, teachers seem reluctant to integrate this information into their teaching practices. However, data use is crucial in a development-oriented system of central tests. Certainly in Flanders (where central tests will be implemented from the school year 2023-2024), but also in an international context, the use of formal performance data to improve student learning can (still) be seen as a change. For change to occur, teachers must be ready to do so. Our research shows that Flemish teachers' readiness is limited: they have rather a limited positive attitude towards data use, consider themselves sufficiently competent, but at the same time they do not expect to have enough available time. Because the use of formal performance data does not occur in isolation, the correlation with school culture was also studied using multigroup SEM. The final path model shows that experiencing a stimulating school culture has a positive impact on teachers' readiness to use formal performance data. In this regard, this study strengthens the bridge to the use of formal performance data from central tests to improve student learning.

## Assessment that is reactive to unforeseen circumstances (e.g. Covid 19) II

15:45 - 16:15

### Hybrid model of high-stakes testing in the Czech Republic: challenges posed by conducting simultaneous online and paper-based exams

L. Firtova[1]

[1]Scio, Czech Republic

Scio, a standardised testing company in the Czech Republic, has been administering the National Comparative Exams to university applicants since 1995, with about 70 thousand exams taken each year. Prior to the pandemic, the exams were exclusively paper-based, but due to the Covid-19 outbreak, Scio has developed an online testing and proctoring solution called ScioLink. After the pandemic, the National Comparative Exams have become hybrid, with students able to choose between paper-based and online version of the test. However, this hybrid format presents unique challenges. One challenge is to ensure that there are no significant psychometric differences between the two test versions. The data have shown that there are very few questions with significant differential item functioning. Another challenge is the prevention of cheating, which is a harder task in online exams compared to paper-based exams. This challenge is addressed through AI algorithms and statistical analyses, which help human proctors detect cheaters. Most of the expelled students are not deliberate cheaters though, they simply fail to follow the rules. The third challenge is the user experience. Scio has conducted surveys to gather student feedback on hybrid testing, with the feedback being mostly positive despite some concerns about comparability and cheating.

# Assessment and Learning during Covid-19 times: perspectives and experiences of university students in Initial Teacher Education in Malta

J. Milton[1], J. Deguara[1], C. Bonello[1], R. Camialleri[1], T. Muscat[1]

[1]University of Malta, Malta

The COVID-19 pandemic disrupted education and Higher Education was no exception as universities replaced face-to-face programmes with remote technologically mediated online courses. This paper aims to explore how early childhood and primary student teachers in Malta experienced learning and assessment during the pandemic through an online on survey.

A number of close-ended questionnaire items explored the modes of teaching and learning student teachers prefer, whether they felt engaged in their learning, and how they were assessed on their course work. Participants were asked whether they experienced a teaching practice placement, how this was assessed and whether they felt supported in their learning. They were also asked whether they would prefer to retain a blended mode in a post-pandemic era.

Results indicate that student teachers experienced various modes of teaching, learning and assessment for their course work and field-practice. Over one third of respondents indicated that they were not given the opportunity to carry out field placement observations. The majority of the respondents prefer retaining a blended approach to teaching and learning, as opposed to exclusively online or face-to-face provision of learning and assessment.

# Co-constructing national qualifications - A novel approach to developing GCSE requirements

O. Stacey[1]

[1]Qualifications Wales, United Kingdom

The principle of co-construction, that is active involvement and collaboration between a variety of stakeholders has been a central feature in the development of the new curriculum in Wales. A number of potential benefits of co-construction have been identified, including increased confidence and buy in from stakeholders in the reforms and greater opportunity for a wider variety of voices to input into the reform process.

A similar co-construction approach has been adopted in the development of the high-level content and assessment requirements for the new set of GCSEs to support the curriculum. This ambitious approach involved a range of participants including teachers, learners, representatives from further and higher education, unions, exam boards and employers.

Results from surveys and interviews with participants in the process identified strengths, challenges and tensions inherent in the process. These will be explored in the presentation, alongside an outline of the final GCSE assessment requirements.

Whilst the benefits of involving a wide range of perspectives in the process clearly strengthened the final requirements, challenges such as a lack of technical assessment expertise among participants and balancing the sometimes polarised views of participants needed careful management.

# Assessment of Practical Skills I

15:45 - 16:15

## Challenges and opportunities in reforming assessment of school principals' capabilities

F. van der Kleij[1], P. Taylor-Guy[1], M. Lasen[1]

[1]Australian Council for Educational Research, Australia

Education systems internationally are grappling with challenges in attracting, retaining, and developing effective school leaders. This presentation draws on evidence from a recent project conducted in partnership with one Australian education system. This project focussed on the development of an evidence-based Principal Capability Framework (PCF) and associated assessment and support resources to guide principal professional growth, with a view to maximising principals' impact on student and staff outcomes. Principal capabilities are future-focused; they include the range of knowledge, skills, ways of thinking and dispositions that enable principals to lead schools successfully, now and in the future.

Through this project, we (1) generated conceptual clarity around the nature of the construct 'principal capabilities'; (2) identified the key capabilities that underpin effective principal practice; and (3) developed a professional growth continuum for each capability to guide self-assessment and self-reflection. These resources may be used along various points of a principal's career, for a range of purposes, including identification and selection of future principals, and supporting ongoing professional growth of experienced principals. Preliminary evidence highlights the value of the PCF, the need for targeted professional learning, and challenges in finding a balance between professional growth and accountability when embedded within wider system initiatives.

# To change or not to change? The case of changing the mode of coursework in Advanced Level Computing to improve its reliability and validity.

A. Grixti[1,2], R. Cini[1,2], S. Mifsud[1,2]

[1]University of Malta, Malta
[2]MATSEC, Malta

Assessment of a subject which includes a coursework component can be complex because of various issues with reliability and validity being of particular concern. In this research, the correlation between the performance of students in coursework and in written examinations in a number of subjects in the science area was explored. A positive correlation resulted. Moreover, in one of the subjects, the mode of coursework was changed since issues of reliability and validity were being flagged from various stakeholders. It was explored whether there was similar or different correlation between the marks of the coursework and the written examination when this change was implemented. When the mode changed to centrally set tasks a better correlation was expected but no significant changes were noticed. Various plots and charts are used to visualize these relations and try to give possible interpretations. Change should never be feared, even when planning and implementing changes in assessment policies, procedures and strategies, however these should always be accompanied by detailed evaluation to maximize its effectiveness. This should then be reflected in a more reliable and valid assessment.

# Psychometrics and Test Development II

15:45 - 16:15

## Exploring the assessment of IELTS writing task response among the high school students

V. Gainanova[1], I. Ismailova[1], X. Alikulova[1]

[1]Nazarbayev Intellectual School of Physics and Mathematics in Shymkent, Kazakhstan

This study reviewed the latest theoretical viewpoints which were relevant to that topic.The study also investigates the challenges the teachers and students face and explores how students can develop their ideas in answering IELTS tasks. The method, used in the current research was a case study. Students' MOCK IELTS writing task 2 was analyzed and the error data were identified. The qualitative research method was employed. The data were collected from a small number of participants ( English teachers) using semi-structured interviews on the perception of students' writing task 2 mistakes and examining the reasons,, and qualitative responses were made. The study findings showed that students managed to answer the tasks. Nevertheless,they had difficulties in applying their vocabulary stock while expressing their ideas. Students also had problems structuring their essays and answering the main questions by using high range of grammar structures. This research has been beneficial both for sides. This study revealed what standards the teachers should set while checking writing task 2 paper in IELTS exam.

# A domain-specific assessment of the critical thinking in universities: from methodology to implementation

E. Orel[1], K. Tarasova[1], D. Gracheva[2], D. Talov[1]

[1]National Research University Higher School of Economics, Russia
[2]Higher School of Economics, Russia

We present the results of assessment of critical thinking (CT) and its progress during the course “Economic Thinking”, which is mandatory for first-year economic students in one of the Russian universities and has CT as a key outcome. Using the Evidence Centered Design, we have developed a CT assessment tool (CT Test) that is based on the economics context.
The theoretical framework includes several components of the CT: 1) verification of information, 2) argumentation and hypotheses, 3) economic analysis, 4) reflection. Two tasks were developed, one of which was administered at the beginning of the course, and the second - at the end. The design of the study made it possible to draw conclusions about the progress of students.
In the study 420 students participated. In our presentation, we will talk about the psychometric properties, the factor structure and feedback, as well as the results of measuring progress in the course.

# Poster session

11:30 - 12:45

## Reduced grading in vocational education

D. Normann[1]

[1]Norwegian University of Science and Technology (NTNU), Norway

Since 2010, increased emphasis on Assessment for Learning (AfL) in national policies in Norway have led to more systematic work on assessment and grading, which seem to have enabled schools to explore reduced grading practice (Directorate of Education, 2019). Some argue that reduced grading enhances motivation to learn for mastery rather than for grades (McMorran et al., 2017); however, little is currently known about reduced grading at all levels of education. The purpose of this study is to investigate how vocational schools implement reduced grading practice and how reduced grading affects vocational training. Focus group interviews with 12 vocational teachers and 9 school leaders was conducted at three vocational schools. Data was analyzed using thematic analysis (Braun & Clarke, 2006). The results show that AfL in national policy enables vocational teachers and schools to implement reduced grading by emphasizing process-oriented learning and teachers' autonomy to make assessment decisions that foster learning. The practice varied between teachers as some chose to ignore school policy by continuing to grade students. The study highlights that vocational teachers' autonomy and decision-making through reduced grading practice directs learning towards vocational competence development rather than performance, which contributes to prepare students for professional practice.

# Accessibility considerations for Digital Assessments - Development of a Framework

S. Mistry[1]

[1]Cambridge Assessment International Education, United Kingdom

The development of classroom-based digital assessments and certified digital high stakes assessments provides an opportunity for Cambridge International to enhance the accessibility of our digital assessment products for as many learners globally as possible. A digital assessment allows for affordances that traditional paper pencil methods may not facilitate.
This poster will summarise the thinking and associated research that has been taking place within Cambridge International on accessibility considerations for digital assessments through the development of a framework. The framework has been designed under five different principles of accessibility research and is primarily for use by assessment designers in the development process. Each principle falls under one of the following strands of research:
- Assessment content,
- Test design,
- User Interaction,
- The use of assistive technology,
- Affordances offered by accessibility tools within platforms.

The poster will also define each accessibility principle with relevant examples. The practical element of the research to date considers a user centred design approach and understanding the specific needs of the test taker, whilst adhering to appropriate digital accessibility standards and guidelines. Consideration is also given to universal design principles and the additional features given by digital assessment platforms to enhance the overall accessibility.

# ‘Disruption' and/or 'Innovation? The case for e-assessment

G. Clark[1], S. Shaw[2, 3]

[1]Scottish Qualifications Authority, United Kingdom
[2]Faculty of Education, University of Cambridge, United Kingdom
[3]Institute of Education, University College London, United Kingdom

In an educational assessment context, there has been a shift towards a learner-centred approach from a system-focused one. This, paired with an “inexorable and inevitable” infusion of technology within all aspects of learning and assessment, affords both opportunity and disruption. On the one hand, the introduction of digital technologies might be thought of as a desirable part of innovation, particularly regarding the new affordances that the technologies may create and the alignment it has with prevailing societal philosophy. Conversely, technology alone cannot transform assessment. Digital innovation can engender disruption in its implementation and use, to deliver improvement in teaching and learning practices and outcomes. For example, moving away from digital conversion of current assessment models and content to full transformation of assessment. This poster seeks to illustrate how the effective use of technology in assessment can disrupt current assessment practice, policies and outcomes, while providing opportunities for better, more valid assessment, focused on the needs of the learner - opening up discussion and discourse to impact positively on assessment practices. It is hoped that this poster will restart and reset the debate over the role of technology in educational assessment.

# Exploring the social and cultural factors that impact on student attainment

P. Surridge[1]

[1]AQA, United Kingdom

This poster presentation draws on existing literature exploring the impact of gender, ethnicity and socioeconomic status on students' attainment within the UK and how these factors intersect to further disadvantage some students.
While these factors may be outside of our control, we are committed to ensuring that our assessments are fair and equitable. We have therefore been conducting extensive work to ensure students from different backgrounds are not disadvantaged by any of our qualifications or any assessment reform we enact.
Students' educational journeys are shaped by their experiences both in and outside the classroom, and research highlights how these experiences and demographic differences may result in differential outcomes. It is therefore crucial when undertaking assessment reform that we consider the wider social and cultural challenges that students may face.
Through our poster presentation we highlight possible reasons for the differential outcomes of students, such as access to resources and support, students' attitudes, gender socialisation and teachers' unconscious bias. We offer some insights into approaches and initiatives that may help to address these social and cultural challenges.

# TALK: Developing a baseline oracy framework for Teaching, Assessment, Learning, and Knowledge (TALK) in school

L. Chvala[1], A. Kaldahl[1]

[1]Oslo Metropolitan University, Norway

Globalization has led to heightened connections of languages and people and increased multilingual realities in society and school. To meet these realities, curricular reform in 2006 designated oracy as a key competence to be developed by all teachers in school (Berge, 2022). Oracy, defined as the ability to speak, listen and interact, was acknowledged as fundamental to human development, literacy development, and eventual professional and democratic participation (Mercer et al., 2017). Since the late 19th century, teachers in Norway have been responsible for the school-based assessment of learners' oracy. This has resulted in well-established, if often tacit, practices in the absence of a central framework for oracy development (Aksnes, 2016; Bøhn, 2015; Chvala, 2020; Kaldahl, 2019, 2020).

This project aims to generate a baseline Knowledge framework for oracy for Teaching, Learning, and Assessment purposes (TALK). The framework aims to bridge the understanding of the research and scholarly community and teachers' established practices. To address a wider and more multilingual understanding of oracy, the framework will focus on two main languages in Norwegian basic education, Norwegian and English. The framework aims to generate baseline descriptions of first and additional language oracy at key stages within the school system.

# The overall impact of cross-language Differential Item Functioning at the test level: The case of PIRLS 2016 in South Africa

H.L. Kayton[1]

[1]University of Oxford, United Kingdom

The test-level comparability of an assessment, particularly when group results vary, is an undervalued consideration. This paper presents an interpretable, thorough and reproducible approach to analysing test-level comparability that can be applied to the evaluation of large-scale assessments and can aid in the development of tests for diverse contexts. Differential Item Functioning (DIF) analysis was applied to investigate the within country comparability of two of the eleven language versions of PIRLS 2016 conducted in South Africa (English and Sepedi). Using an IRT-based likelihood-ratio test approach to DIF detection, findings show that when a suitable anchor was used, almost a quarter of items function differently across versions. The impact of DIF at a test level was then explored by considering four indices of Differential Test Functioning (DTF) that consider the direction, type and effect size of the DIF found. DTF effect size statistics show that the overall effect of DIF was significant and impacted the comparability of the test across language groups. Moreover, the extent of differences in achievement across groups is being obscured by differential functioning of the test; despite being the lowest performing group, the DTF indices show that Sepedi students had an advantage at the overall test level.

# IB Open Book Exams pilot study – A picture of our schools at the start

R. Chivers[1], R. Hamer[1]

[1]International Baccalaureate, Netherlands

While open book exams (OBE) are common to reap the benefits on student wellbeing and learning outcomes, as well as positive backwash on teaching, it needs to be implemented in the right way. However, the literature on what this “right way” looks like is fragmented and does not cover international high school level education. As of May 2023, the International Baccalaureate (IB) has recruited 332 schools worldwide, of which half will pilot three OBE formats that the IB is not yet offering. The other 166 control schools will offer the course and exams as agreed in the current course. In recruiting the pilot and control schools were carefully matched by region, language of instruction, school type and cohort sizes. This was to ensure that the school groups would be very similar and the impact of the educational reform would be measurable as the difference between the groups. This poster will present a check on the study design and recruitment outcomes by examining the similarity across school groups, where the baseline survey data allows a comparison of their student and teacher populations. Results presented will comparison of teaching strategies, wellbeing, growth mindset and learning styles at the start of the pilot.

# Is that true? No, that's nonsense! Understanding AI hallucination and confabulations

R. Hamer[1]

[1]International Baccalaureate, Netherlands

The wonderment of many users of powerful generative AI has not abated much since its release. There are many media posts and articles introducing new ways of using the large language models (LLM) to write messages, daily schedules, summaries and so on. At the same time, the initial extreme concern voiced by researchers and educators regarding the demise of human writing and creativity has somewhat abated because the limitations of LLM are being recognized, at least by knowledgeable users.

One such limitation of LLM is producing so-called AI hallucinations. AI hallucination may better be defined as confabulations: where highly probable and potentially accurate information is combined in ways that form sentences and meanings that prove to be false. While more recent models claim to have addressed this issue, this poster will share results from two small studies demonstrating the quality of output and the frequency and character of AI confabulations across three languages. Differing levels of AI confabulation across tasks and languages is the logical outcome of the very nature of LLM and the limitations of the training corpus. Understanding how these impact AI output quality and validity for different purposes is core to AI literacy.

# Evaluating and assessing distance education learners: Developing a comprehensive learner model

S. Radović[1], N. Siedel[1]

[1]Center of Advanced Technology for Assisted Learning and Predictive Analytics, Fern University, Germany

Technological and pedagogical advances are redefining education. The integration of technology in distance higher education has opened new opportunities for personalized learning experiences, facilitating teaching methods that cater to learners' individual preferences, skills, and goals. To achieve this, a comprehensive understanding and evaluation of learners' learning progress and success is necessary. A learner model, which is an abstract representation of a learner based on the digital traces left in the learning environment, is often used for assessment and evaluation purposes. This paper provides a short overview of existing models and proposes a new learner model. The proposed model identifies a comprehensive set of parameters necessary for assessing different aspects of the learning process, including students’ demographic, previous knowledge, professional skill, and experience, as well as cognitive, meta-cognitive and social activity data. While there are various alternative learner models in the literature, the proposed holistic model is advantageous in terms of comprehensiveness, as it can be used to evaluate learning progress and success from several perspectives. Our approach suggests that the combination of such various perspectives on learners' data is not only appropriate for assessing students in higher education but also a useful approach for initializing learning personalization and adaptation.

# Read Messick! Developing ethical AI will require assessment literacy.

C. Aloisi[1]

[1]AQA, United Kingdom

This poster presentation argues that AI development has 'validity' issues that could turn into assessment validity concerns were these systems deployed in a high-stakes context.

Explainability is a well-known limitation of and barrier to ethical AI. While solutions are being developed, it is still unclear what state-of-the-art AI can and cannot do. This is not just an algorithmic ‘black box’ issue; indeed, an under-explored research area (particularly outside of the computer science community) is how AI systems are trained, given feedback and evaluated by people.

Preliminary work suggests that the formal assessment of AI capabilities is done, even by leading companies, by non-assessment experts. The people in charge of evaluating AI performance on specific tasks are often not experts in those tasks. Likewise, the tests AI systems are subjected to are not 'tests' in the psychological or educational measurement sense, but they are similarly used to make inferences about an AI’s 'knowledge and skills'.

In other words, we argue that the validity of the claims around AI capabilities sometimes relies on tenuous evidence. We hypothesise that this may be to lack of assessment literacy by computer science experts. We propose research to test this hypothesis and suggest tentative solutions.

# A meta-analysis of math anxiety interventions

E. Sammallahti[1], J. Finell[2], B. Jonsson[2], J. Korhonen[1]

[1]Åbo Akademi University, Finland
[2]Umeå University, Sweden

The experience of math anxiety can have detrimental effects on students' math performance, and researchers have in recent years tried to design interventions aiming at reducing math anxiety. Our meta-analysis aimed to examine the effectiveness of math anxiety interventions in reducing math anxiety and improving math performance. The meta-analysis comprised of 50 studies and included 75 effect sizes. On average, the effect sizes were moderate (g = -0.467) for reducing math anxiety and improving math performance (g = 0.502). Furthermore, our results indicated that interventions that focused on Cognitive support or regulating Emotions were effective both in reducing math anxiety and improving math performance. In addition, longer interventions and interventions targeting students older than 12 had the biggest decrease in math anxiety. To examine study quality, we utilized the Quality Assessment Tool for Quantitative Studies (EPHPP) to classify every article into one of three categories of quality: 1) High, 2) Medium, and 3) Low. Study quality was however not related to intervention outcomes.

# Not like that! Attempting to use GPT to generate examples in statistics

I. Casebourne[1]

[1]Digital Education Futures Initiative (DEFI), The Bridge, Hughes Hall, University of Cambridge, United Kingdom

Generative AI is currently being used to generate items that can then be checked and validated by humans (for example, Duolingo DET). Inspired by Brigg's suggestions about learning progressions, a potential use for AI might be to help teachers to generate formative test items or graded worked examples which might assist students to construct models or schema. Students may encounter one or two exemplars, but not examples of poorer performance or errors. Without encountering and understanding poor examples, it may be more difficult to understand what is good. However, there are currently a variety of barriers to presenting students with such examples. Generative AI has the potential to help, providing responses to a question at various grades with commentary explaining how poorer scoring examples might be improved. This poster presents the author's attempts to generate examples for statistics, first using ChatGPT, then Bing, then GPT 4. It discusses the limitations of large language models for this purpose, such as basic maths errors.

# Construct definition in international educational assessment design

L. Badham[1]

[1]International Baccalaureate, United Kingdom

Creating fair and valid assessments for linguistically and culturally diverse cohorts is fraught with challenges. Construct bias, where groups of students understand or represent a target construct differently, poses a particular threat to validity and score comparability in international education and assessments. Clearly defined constructs can support the development of fairer and more valid assessments, yet literature on the approaches and processes of construct definition in assessment is surprisingly scarce.
This poster will present ongoing doctoral research from the University of Oxford and the International Baccalaureate (IB) exploring different approaches to construct definition in international assessment design, and methods will be presented for interrogating target constructs to reduce unfairness and bias. A review of academic and grey literature establishes the state of the field regarding construct definition, and a case study method will be used to examine how constructs are defined and developed in practice in different assessment contexts. Initial results will be presented for IB curriculum-based assessments and PISA's psychometric approach and, if possible, these will be compared to an outcomes-based assessment model.

# Assessing the Swedish Shortened Mathematics Anxiety Rating Scale and its Relationship to Math Performance and Attitudes in Young Students

J. Finell[1]

[1]Umeå University, Sweden

The current study assesses the Swedish shortened version of the Mathematics Anxiety Rating Scale – Elementary (MARS-E). The original scale included 36 items and was conceptualized as a two-factor construct including a cognitive and an affective component of math anxiety, specifically respondent's "worry" and "nervousness" respectively. Building upon previous theory from Henschel and Roick (2017), this study examines the dimensionality of the instrument and its invariance across gender and time.

The reliability of the scale is assessed and compared to the original version developed by Henschel and Roick (2017). In addition, this study also evaluates relationships between the MARS-E and other criterion variables. These include math performance, test anxiety and math self-concept. Correlation analyses and structural equational modelling are employed for this evaluation. Effect sizes obtained from these analyses are compared to previously reported findings.

The findings from this study will confirm or falsify the robustness and dimensionality of the MARSE in the context of a young Swedish population. Furthermore, conclusions of relationships with other variables will contribute to the existing body of research and enhance our understanding of the development of math anxiety in young individuals.

# Feedback Culture at School: What Remains Neglected?

Z. Utesheva[1], S. Unbayeva[1]

[1]Nazarbayev Intellectual School for Chemistry and Biology, Aktau, Kazakhstan

It is not new in the field of education that feedback between teacher and student is one of the main parts of effective learning and teaching. However, when it comes to practice, it is an important skill that requires the most experience and competence from both sides during the learning process. This poster shares the results of research on the factors and obstacles that affect the rationality of the feedback exchange process in the student-teacher environment, including the feedback giver, the feedback receiver, the environment in which the feedback takes place, as well as its content, accuracy and plan-based characteristics.
The results of the study revealed that both students and teachers understand the importance of the role of feedback in the educational process, but in reality there are barriers such as time pressure, lack of competence and formality.
The main conclusion of the research is that in order for learning to be most useful, both sides of feedback should have certain qualities, these qualities can be developed through special trainings and practice, and for this, a comfortable, safe, supportive feedback culture should be created in the school.

# Does a unitised approach build resilience into an assessment system?

R. Harry[1]

[1]WJEC, United Kingdom

Following COVID-19, assessing some parts, or units, of a qualification midway through a course of study is often presented as increasing assessment system resilience, by providing a basis for determining fair, reliable and valid grades if examinations are cancelled.

In Wales, A levels are unitised; 40% of content is assessed in ‘AS units’ in the first year of study, with the remaining 60% assessed via terminal assessments at the end of the second year. The aim of this study was to establish whether this design feature provided resilience when terminal A level assessments were cancelled in 2020.

Although performance in AS units was a good predictor of a student’s final A level grade, modelling shows that several grade outcomes remained plausible for students. Faced with uncertainty over a student’s most likely outcome at the end of the course, teachers tended to award plausible, but positive, grades. Cohort outcomes were as inflated as A level results in another country, England, where no equivalent information was available to support grade judgements.

The findings show that a unitised approach is not sufficient to cohere student and cohort level definitions of standards when exams are cancelled, and thus cannot provide resilience in such circumstances.

# ‘But what do we do with the results?’ A systematic approach to using assessment data holistically to improve teaching and learning

S. Crocker[1], I. Suto[1, 2]

[1]Cambridge University Press and Assessment, United Kingdom
[2]Cambridge CEM, United Kingdom

Background
When defining educational success, many cultures include life competencies, moral values, and high wellbeing and alongside academic achievements and progression to further education or employment. Teachers can assess or evaluate all these elements of holistic education. However, not all are confident in combining different types of data to support teaching and learning.
Methods
Drawing on academic literature and classroom experience, we developed an iterative six-step approach to collating assessment data, both quantitative and qualitative, and using it formatively. The steps are: Map, Assess, Analyse, Plan, Shape teaching and learning, and Reflect and refine.
Working with 80 international schools in Malaysia, Indonesia, Thailand and Vietnam, we introduced the approach to 200 teachers in face-to-face workshops. Working in small groups, the teachers identified the data types they currently collected and reviewed their own practices against the approach. Afterwards, they gave feedback through individual questionnaires and plenary discussions.
Findings and discussion
Almost all teachers valued the approach, reporting it helped with organising and using data more effectively. However, whilst there was a widespread desire to adopt a holistic approach in understanding and supporting individual student needs, in practice limited resources and teacher experience could compromise its comprehensiveness.

# Cooperation in external assessment – projects in Cape Verde and Angola

A. Monteiro[1], M. Gomes[1], M. Borges[1]

[1]Instituto de Avaliação Educativa, IAVE, Portugal

IAVE has been establishing cooperation protocols with Cape Verde and Angola, members of the Community of Portuguese Language Countries (CPLP). These protocols aim at developing projects to strengthen the implementation of national systems for the external assessment of learning, thus promoting “Quality Education” (SDG4, 2030 Agenda).

IAVE shares its expertise with these countries’ Education Ministries in human resource training and technical support in:
• design, construction and validation of external assessment tests;
• marking and marking supervision;
• design of technological tools for process management;
• production of statistical and qualitative reports on results.

Cape Verde
• 2019 – Low-stakes Testing I: Portuguese and Mathematics, grades 2 and 6 (sample nationwide);
• 2023 – Low-stakes Testing II: Portuguese and Mathematics, grades 2 and 6 (sample nationwide).

Angola
• 2022 – National Exams Field Trial: Portuguese and Mathematics, grades 6 and 12 (sample nationwide);
• 2023 – National Exams: Portuguese and Mathematics, grades 6, 9 and 12 (extended sample nationwide); Natural Sciences, grade 6, and Physics, grades 9 and 12 (sample nationwide).

Cooperation has contributed to improving the validity and reliability of external assessment, with impact on the quality of these countries’ educational systems, as well as their pedagogical practices.

# Development of a framework for assessing mathematical literacy in primary and secondary school: A pilot study

M. Mikite[1], I. France[1], Ģ. Burgmanis[1]

[1]University of Latvia, Latvia

Mathematical literacy determines qualification of school graduates. Referring to PISA international comparative study Latvian students underperform in high-level mathematical literacy tasks. Results above the OECD average is one of the targets in the National Development Plan of Latvia. Learning objectives in Latvian curriculum of mathematics are closely aligned to those measured by PISA. To keep track of progress towards goals, the authors are developing a framework for assessing mathematical literacy in primary and secondary school. This assessment tool aims to compare results across schools and over the years. The dimensions of the framework are (1) the level of mathematical reasoning, (2) the content of mathematics, (3) proficiency level of mathematical literacy learning. Assessment tool tend to be age-independent. It allows tracking the process of mathematical literacy development and results can be used by teachers. In this pilot study, the authors created a test, and it was piloted by 93 students in three different schools. Results were analysed using a IRT Rasch model. The results show a gap between students' performance in middle cognitive level and high level tasks. The results highlight the gap between the standard requirements and students' performance and suggest direction in which to improve the teaching approach.

# Divergent considerations during the journey to internationalise mathematics questions in an adaptive baseline assessment.

E. Barthel[1, 2], I. Suto[1, 2]

[1]Cambridge CEM, United Kingdom
[2]Cambridge University Press and Assessment, United Kingdom

Background

Many UK schools use standardised baseline assessments each academic year, valuing the year-on-year comparability of results data. We explore a tension arising in our journey to reform a long-established computer-adaptive baseline assessment. With an increasingly international user-group, we sought to improve accessibility whilst maximising the comparability of outcomes pre- and post-reform.

Methods

350 mathematics questions were reviewed systematically by experts in mathematics, accessibility, and linguistics. This included evaluating language demands against the Common European Framework of Reference and appraising the inter-cultural recognisability of images. Potential amendments were identified, and probable impacts on item difficulty and validity were hypothesised and ranked.

A subset of items was adapted, and the amended assessment was trialled with children with a range of abilities in diverse schools internationally. The trial data was analysed to ensure continuity of item level difficulty had been maintained and to highlight any mismatches.

Findings and discussion

Encouragingly, the reform journey revealed item amendments engendering the greatest accessibility improvements are not always the hardest to make. Stakeholders varied in the relative values they placed upon accessibility and comparability of results. This poses a challenge for assessment developers in determining what to prioritise.

# Student motivation in history: associations between formative assessment, historical consciousness and ‘doing history’

H. Eriksen[1], H. Roaldset[2], K. Korbøl[2]

[1]Oslo Metropolitan University, Norway
[2]University of Oslo, Norway

History curricula in Norway has undergone change through reforms, the latest from 2020, and the assessment system has been rewritten starting in 2007 which gave formative assessment a legislative status which holds until today, which is special in an international context. However, the knowledge about how formative assessment is perceived by students in different school subjects and within important aspects of history education is limited. The aim of this study is to explore the associations between students’ perceptions of historical consciousness, historical inquiry (‘doing history’), formative assessment and students’ motivation for learning history. Data were collected through a nationally distributed survey (N = 569) among students with a broad experience from history education in upper secondary schools in Norway. Given the unexplored status of the field with the current methods, we conducted exploratory and confirmatory factor analysis to investigate the operationalisation of the latent variables. Further, using structural equation modelling, we estimated the strength of the connections between the variables. Findings indicate significant associations between student motivation as the dependent variable, and the independent variables of historical consciousness, ‘doing history’ and formative assessment. The study has consequences for practitioners, researchers, and policy makers.

# Comparative Judgment vs. Criteria-based Assessment in Legal Education

K. Egelandsdal[1], E. Hartell[2]

[1]University of Bergen, Norway
[2]KTH Royal Institute of Technology, Sweden

The poster will present findings on the formative use of comparative judgment and criteria-based assessment in a course on administrative law as part of a master's program in legal education. Students used the two assessment approaches to assess the quality of administrative decisions. A student population of 300 was divided into two cohorts and given six administrative decisions of varying quality to assess. One cohort assessed the decisions using a comparative judgment approach, and the other used a criteria-based assessment. In both cases, the students assessed the legal decisions focusing on a) method, b) language, and c) content and ranked them from best to worst. Before the assessment activity, the students had practiced writing administrative decisions on their own and participated in a lecture that prepared them for the assessment activities. After the assessment activity, the students discussed the quality of the decisions they had assessed in groups with participants from both cohorts. The focus of the study was to investigate whether and how the students experienced the two assessment approaches contributed to developing their understanding of quality in administrative decisions. Data was collected through a student survey and the student ranking of the legal decisions.

# Comparative Judgment for Summative Assessment in Legal Education

K. Egelandsdal[1], E. Hartell[2]

[1]University of Bergen, Norway
[2]KTH Royal Institute of Technology, Sweden

The poster will present findings from a study on the use of adaptive comparative judgment by examiners as a method for assessing, ranking, and grading student exams. The study aims to investigate if and how this method may support examiners in making better judgments when assessing student works for grading. In this study, approximately 300 essays will be assessed by 20 examiners using adaptive comparative judgment. These essays have already been assessed and graded using a traditional criteria-based approach with one examiner per 30 essays. These grades will serve as a point of reference for the study. The examiners using comparative judgment will use the same assessment criteria as in the traditional approach, but the essays will be assessed in pairs. In this pairwise comparison, the examiners chose which essay is better and justify their choice based on assessment criteria for each pair. Using learning analytics, the judgment of all essays will be ranked in terms of quality. 2 professors will read through the ranking and try to identify the benchmarks for the lowest quality work for each grade. The benchmarking will be based on the professor’s assessment of quality considering the assessment criteria, not on a normal distribution.

# Transformations in Large-scale Educational Assessments: The Case of India Compared Internationally

P. van Rijn[1], I. Bhaduri[2], J. Bertling[3], H. Por[3]

[1]ETS Global, Netherlands
[2]National Council of Educational Research and Training, India
[3]Educational Testing Service, USA

Assessment reforms are often informed by the changing needs of society. In recent years, India has undergone significant assessment reforms to improve the quality and relevance of assessments. One key measure of student achievement in India is the National Achievement Survey (NAS), which is a national-level large-scale educational assessment in grades 3, 5, 8, and 10. The NAS 2021 was built on previous national surveys and expanded to provide a comprehensive assessment of learning outcomes with increased content coverage. With India being one of the largest countries in the world with state-run public education with differences in student achievement between states, comparisons with the design and implementation of other, international large-scale educational assessments can be helpful to provide an outlook for transforming national assessment in India and beyond. In this poster, we compare the NAS with the Programme for International Student Assessment (PISA), to provide insights into the similarities and differences in terms of their objectives, assessment framework, sampling and measurement designs, analysis methodology, and approaches to reporting. The findings of this study will contribute to the ongoing discourse on assessment reform and will inform future education policies and practices in India.

# Speak properly! Understanding the role of auto-generated captioning technologies in the marginalisation of disabled speech

C. Tupling[1]

[1]AQA Education, United Kingdom

Automatic speech recognition technologies are increasingly used in educational contexts to support learning, teaching and assessment. They are often positioned as labour saving and inclusive. Relying on AI to convert speech audio into text they have a high accuracy in generating automated captions from normative speech patterns. Less accuracy is observed in the captioning of disabled speech and thus these technologies potentially represent ableist forms of exclusion.

This poster shares results from the author's autoethnographic study examining automated captions generated using a popular lecture capture and speech recognition platform used in UK Higher Education. As a person who stammers (pws) the author was in a unique and privileged position to both experience and critically interrogate the inaccuracy of captions generated and the poster offers an insight into the problematic ways AI converts stammered speech into text.

The ways in which speech recognition technologies rely on biased training sets is considered and examples of how this leads to the marginalisation of disabled speech through mis-transcription is evidenced. Developments in more equitable forms of speech recognition technology are considered and proposed in order to progress their socially just use in educational assessment.

# An investigation of approaches to student assessment in international high schools in China in the context of practices internationally.

X. Yang[1]

[1]Trinity College Dublin, Ireland

Rationale:
Between 2013 and 2022 international schools in China increased from 505 to 949. Anecdotal evidence suggests that achievement in these schools is high. However, little is known about how schools' internal assessment practices are planned and implemented and how such practices might be associated with achievement.

This poster presents results of a study identifying common characteristics of assessment approaches in International Schools globally. Exploring how International Schools in China manage the tension between summative and formative assessment will help illustrate the critical importance of context in framing schools' assessment policies.

Research Question 1. What are the main characteristics of approaches to assessment in international schools globally, including the balance between formative and summative purposes?

Research Question 2. What are similarities and difference in assessment in international schools in China compared to other global international schools?

Research Question 3. What are the main characteristics of approaches to assessment of chemistry in international schools in China?

The study employs a systematic literature review combined with a review of websites and assessment policies in selected schools in China.

Findings will contribute to a better understanding of assessment practices, opportunities and constraints for International Schools in China, associated with assessment cultures.

# Towards justified use of automated speaking assessment algorithms via an argument-based validation: A case study of prosodic features assessment

Y. Hao[1]

[1]University of Oxford, United Kingdom

Reforms that call for an increasing use of automated language assessment have been fuelled by both stakeholders' needs for individualised, flexible yet valid assessment methods especially after the Covid, and the remarkable improvement in performance of algorithms such as neural networks and large language models (e.g., ChatGPT). Some of these algorithms are incorporated in large-scale language tests such as TOEFL and Duolingo. Despite their popularity, there is limited validation research on the use of algorithms in automated language assessment (ALA). Therefore, this study aims to propose an argument-based framework for validating the algorithms used in ALA, with a particular focus on prosodic features in spoken English because assessing these features involves much more complicated processes than assessing other features like vocabulary and grammar. This proposed framework is based on the results of a systematic review of 45 research articles on automatic assessment of prosodic features. The framework consists of the claims, warrants, and backing (evidence) for each of the following inferences: input data, speech recognition algorithms, feature selection, modelling algorithms, and output features used to train the algorithms. This study echoes this year's theme by attempting to evaluate the success of automated speaking assessment via the proposed argument-based validation framework.

# Self-Assessment in performance. Teachers' thoughts and concerns.

D. Tsalta[1], T. Rousoulioti[2], A. Ventouris[2], O. Blatsioti[1]

[1]University of Nicosia, Cyprus
[2]Aristotle University of Thessaloniki, Greece

The present research examines teachers’ believes related to the application of self-assessment in the context of teaching Greek as L2. The aim of the study is to explore the believes of 124 teachers, who teach at Intercultural schools in Greece, about the use of self-assessment in order to assess multilingual students who learn Greek as L2.
The main research questions are:
1. Does the use of self-assessment improve students' performance (Panadero et al. 2017)?
2. Does self assessment lead to students’ personal development (elimination of competition, boost of self-esteem (Rolheiser & Ross (2001))?
Data were collected from an electronic questionnaire and a number of interviews (method triangulation). Research results showed that the majority of teachers develop a positive attitude towards self-assessment and recognize its multiple benefits. Self-assessment serves contemporary student-centered pedagogical orientation and autonomous learning. Self assessment also enhances students’ performance - including some latent aspects of learning motivation and metacognitive skills. However, teachers continue to raise questions on the practicality, validity, reliability and usefulness of this method.

# Assessment that is reactive to unforeseen circumstances (e.g. Covid 19) III

9:00 - 9:30

## COVID-19-related changes to upper secondary assessments in six countries: Adaptations and reactions

N. Rushton[1], S. Lestari[1]

[1]Cambridge University Press & Assessment, United Kingdom

When the COVID-19 pandemic started in 2020, many schools around the world closed and the examinations that were due to take place at the end of upper secondary education had to be cancelled, postponed or altered in some way. COVID-19 continued to affect the assessments in 2021 and 2022 in some countries.
We compared the changes to examinations in six countries (China, England, India, Italy, Spain and the USA), the effect on standards, and the reaction to the changes. Such changes could be expected to generate comment from a range of stakeholders such as teaching unions and the public. It can be difficult to find research evidence citing opinions at the time of changes, but newspapers and other media sources are able to publish such information with immediacy and often report the perspectives of important stakeholders. Therefore, we examined newspapers from the affected countries to identify issues or problems that arose from the changes and opinions about them.
We discovered that the press perceived that standards changed in some countries. However, concerns were raised in every country, although their extent varied. Some, such as increased anxiety, affected multiple countries, whilst others, such as technological problems, were confined to particular countries.

# How the pandemic impacted on how Scotland and other jurisdictions assess young people and the implications for the future

S. Hill[1]

[1]SQA, United Kingdom

This paper, detailing findings of research into how school qualification assessment and certification were undertaken in 16 jurisdictions in response to COVID-19, puts Scotland's response in a wider context. This research helps us understand whether there are any links between particular national approaches to summative assessment and how effectively those systems functioned during the pandemic. It also provides an overview of the continuing or lingering effects of the pandemic on assessment and certification systems and explores whether the experience is likely to lead to changes to qualifications and assessment in the longer term.

Key findings include:
-Measures taken in different jurisdictions were driven not just by practicability but by individual jurisdictions' societal attitudes towards assessment.
-While the early experiences of the pandemic may have exposed weaknesses in a reliance on exams, measures put in place since may have exposed the limitations of the alternatives.
-There are increasing questions about whether the traditional concept of fairness in assessment is wide enough.
-There is little consensus on what the long-term effects of the pandemic on assessment should or will be, but the experience has re-focused discussions on the purposes of assessment.
-The experience will feed into reforms taking place across several jurisdictions.

# Assessment of International GCSE English: Insights from 2 years of live delivery of onscreen exams and implications for regulated environments

A. Ulicheva[1], I. Custodio[1], H. Dalton[1], M. Reeve[1]

[1]Pearson, United Kingdom

In England, we are yet to see a significant transition from paper-based to digital assessments for school-based national assessments for General Certificate of Secondary Education (GCSE). To corroborate the validity of onscreen assessments, a comprehensive roadmap for research, design, and implementation is required. This will help ensure that assessments are fit for purpose, and that any innovations continue to retain stakeholder confidence while complying with regulatory requirements.

This paper shares insights gleaned from a successful attempt to establish a digital alternative to a paper-based test in the context of such a regulated assessment environment. We report on the first two years of live delivery of International GCSEs in English Language and English Literature in schools across Europe and the Middle East.

With consideration given to barriers identified in 2020, we focus on student experiences and voices expressing attitudes and preferences towards onscreen assessment following both practice and live test sessions. We emphasise specific issues around onscreen item presentation and explore links to best practice in onscreen test design. We also discuss schools' familiarity with onscreen assessments and whether it impacts on student experience and performance.

# E-Assessment III

9:00 - 9:30

## Differences in mathematics results due to item types

E. Aldenius[1], J. Ingmarsdotter Lundmark[1], V. Severyd[1]

[1]Stockholm University, Sweden

This study compared responses to ten mathematics items administered with paper-and-pencil to nearly identical items in a digital format. Responses were from 130 students in year 3 (9–10 years old) from 22 different schools located across Sweden. A convenient sample was made with regards to school geographical distribution, urban/rural location, size, and public/private status. The ten items differed according to mathematical content and required varying amounts of writing/typing to show working out. Digital items were categorized based on the type of interaction that occurs between the student and the digital device. Results showed that there was no significant difference in score when only point and click (e.g., multiple choice) or moving an object (e.g., matching or drag-n-drop) was required. However, for items requiring a constructed response, the result differed due to the amount of text entry. When the students only had to type/write a short answer, students' scores were higher in the digital format. When the students had to show their workings, student' scores were lower when responding digitally. Studying differences in performance student by student revealed that this may be due to differing perceptual demands and a need for familiarity with digital devices.

## Summative Assessment I

9:00 - 9:30

### Exploring marking time and examiner agreement for item-level versus whole response marking

A. Furlong[1], L. Badham[2], M. Wami[1]

[1]International Baccalaureate, Netherlands
[2]International Baccalaureate, United Kingdom

In 2009, the International Baccalaureate (IB) began its journey to an onscreen method of marking. Part of this was to move towards marking examination papers at an individual question level (also known as item level marking, segmentation in marking or marking by question item groups, hereafter referred to as ‘QIGs’) for some assessment components.
Whilst the expectation was that this would result in more reliable and quicker marking, there has been uncertainty internally that these benefits are being realised. Therefore, the IB conducted an experimental study in 2021 investigating the marking time, reliability and examiner experience of whole response marking compared to marking by QIG. A group of experienced Physics examiners were asked to mark scripts where half were presented as QIGs and half as whole responses and their marks were compared to the definitive standard. The examiners also completed a questionnaire to provide feedback on their experiences. The results suggest that marking by QIG was not significantly more reliable for this assessment, but it was faster and examiners expressed a strong preference for it. The questionnaire results also suggest other advantages and disadvantages, such as the logistical benefits of being able to move between questions within the same paper.

# ‘How do you assess that?’ Achieving meaningful engagement with a large dataset as part of a reformed A-level Mathematics.

B. Redmond[1], J. Golding[2], G. Grima[3]

[1]Pearson, United Kingdom
[2]University College London Institute of Education, United Kingdom
[3]Pearson UK, United Kingdom

Drawing on a four-year study (2017-2021) which explored reformed pre-university A-level Mathematics in England, we outline the extent to which students have been able to meaningfully engage with a large dataset-focused component of this qualification, and ways in which enactment of the specification, as well as approaches to assessment, have influenced this. The implication of shifts in practice during Covid are also considered.

Mathematics/Further Mathematics A-levels, the calculus-rich pre-university qualifications in England, were reformed for first teaching from September 2017. They include greater content-related scope, and enhanced expectations for mathematical proof, problem solving and modelling. They are terminally assessed by timed written papers. As an aspect of modelling, Mathematics A-level students are required to engage with a large dataset using suitable technology. However, curriculum pressures, as well as limited assessment reward for large dataset work, mean that in study classrooms this element was often sidelined. Access to technology, teachers’ digital pedagogy and students’ IT skills, were also barriers. Pressures associated with covid, and its ongoing impact, then led to further marginalization of the large dataset. Post-covid we saw changes in teachers’ practice and attitudes towards assessment that could offer revitalised approaches for engaging with the large dataset in the future.

# To err is human: how AI might contribute to trust and accountability

D. West[1]

[1]AQA, United Kingdom

Can a classifier based on Artificial Intelligence (AI) address learners' concerns about biased markers? AI can provide marking that is reproducible from day to day. However, AI will not always provide a valid outcome, therefore automated systems need expert human oversight when making high-stakes decisions.
High-stakes assessments often use item scores, summed to a total mark that can be graded. Within each component there is a local valuation of 'a mark's worth' that is refined by experts during awarding. It is not always easy for an onlooker to understand a decision for one item, if they haven't attended a standardisation event.
In trials AI scores matched expert human scores for 84% of open responses, making an AI second score a potent quality control. Instances of scoring disputes between expert and AI provide a washback effect for writing more reliable assessment items.
Empirically we find that within a set of responses those with similar themes can be clustered together, along with their scores, so that any score selection outliers can be scrutinised. Those response themes leading to poor score similarity within a cluster can be investigated further to improve item and mark scheme designs.

# National Tests & Examinations II

9:00 - 9:30

## How are GCSE grades used in post-16 admissions decisions?

E. Walland[1], T. Leech[1, 2]

[1]Cambridge University Press & Assessment, United Kingdom
[2]OCR, United Kingdom

Many recent policy proposals for the future of assessment in England suggest reform of GCSEs, qualifications taken by learners aged 14-16. Since GCSEs are no longer the main educational-exit qualification (as learners must stay in education until 18) some writers have questioned their necessity. Some argue for the abolition of high stakes external assessment in this phase, others for streamlining such assessment. To inform this debate, evidence of how GCSEs are used is needed. We investigate use of GCSE grades in admission processes for post-16 education, including sixth form and further education colleges. Using analysis of documents, questionnaires and in-depth interviews of teachers representing a range of school types and subjects, we explore teachers' views and current uses of GCSE grades in admissions, and what reform of GCSEs might mean for this decision-making. We found GCSEs play the central role in post-16 admissions decisions, and that external national examinations taken at age 16 were valued as selection tools by teachers, though many said they could still make selection decisions using fewer GCSEs. We explore impacts of the current system and any potential changes. We discuss findings in relation to theories of the uses and purposes, impact and consequences of assessments.

# Vocational Assessment in Secondary Education – Recent Developments

J. Muscat[1], R. Cuschieri[2], S. Sammut[2]

[1]University of Malta - MATSEC, Malta
[2]University of Malta, MATSEC, Malta

In 2019, the MATSEC Examinations Board launched eighteen new vocational assessment syllabi in line with the national education strategy encompassed within 'My Journey'. Syllabi for the existing six subjects were restructured while twelve new subjects were introduced. Policy Documents regulating these programmes' assessment were also revised to reflect such changes.

The new syllabi were the first of their kind to be introduced in Malta since they incorporated three different attainment levels within each subject. These levels were modelled on the Malta Qualifications Framework (MQF) levels as proposed by the Malta Further and Higher Education Authority (formerly NCFHE) in its 2016 Referencing Report. In view of the wider range of skills assessed through an increase in assessment criteria targeting different levels, a new Glossary of Terms was published by MATSEC to support assessors in correctly interpreting active verbs.

Such developments posed various challenges related to assessment. Besides, the number of mitigation measures that had to be adopted during the COVID-19 pandemic has impacted initial cohorts. Additionally, the recent restructuring of the Learning Outcomes model for other SEC subjects will potentially influence vocational subjects too. This study investigates some of these situations, shedding light on the way these programmes are advancing.

# Exploring teacher perspectives on assessment reform: the change from modular to linear A-level assessment

G. O'Brien[1]

[1]AlphaPlus, United Kingdom

This research explores responses to reform at a school level. A 2015 educational reform in England to post-16 qualifications was the decoupling of the AS and A level qualifications. These high-stakes qualifications have multiple purposes, including selection for higher education. This reform represented a shift from a modular to a linear examination structure and teachers' perceptions of this change were explored. The findings highlighted the tension between the selection and qualification purposes of A level qualifications and showed that teachers' perceptions of the reform were shaped by contextual factors within their institutions, as well as by the high-stakes nature of these qualifications.

# Higher Education & Assessment

9:00 - 9:30

## Developing a New English Placement Test for Higher Education in Israel: A Survey of Stakeholders

L. Atzmon[1], R. Fortus[1], T. Karelitz[1], H. Lerman[1]

[1]National Institute for Testing and Evaluation, Israel

A concern accompanying the development of a new test is what prospective stakeholders will think of it. Is the test necessary? Are the question types relevant? Is the proposed use of the scores controversial? Gathering information on stakeholder perceptions at an early stage in the life of a test is useful. If there are large gaps between the perceptions of the test developers and those of the stakeholders, the validity of the test could be adversely affected.
This survey examined the perceptions of different stakeholders in a new English placement test for higher education developed by NITE. The goal was to determine whether such gaps exist and – if they do – consider ways of adapting the test to eliminate or reduce them. In the context of learning English at an institute of higher education, we asked what language abilities people considered important, and whether and how an English placement test and its scores should be used.
The survey included interviews with key figures in Israeli institutes of higher education and an online questionnaire in three languages, to which 738 academic staff and 7,296 students from various colleges and universities responded.
The presentation discusses the main findings of the survey.

# Data literacy for educators: A model for transforming data and information into instructional knowledge and practice.

V. Glickman[1], T. Milford[1], J. Anderson[1]

[1]University of Victoria, Canada

Educator Data Literacy is a critical component of effective data use in education. Research shows that when teachers receive effective training and support in using student level data to inform instruction, it can lead to improved student achievement. There are also a variety of obstacles that pre-service and classroom teachers face in using student level data - lack of training, time constraints, privacy policies, technology barriers and resistance to change. Teacher education programs in British Columbia (BC) are required to include classes on how to best assess student learning towards mandated curriculum; however, data literacy – how teachers transform information into instructional knowledge and practice – is not typically a part of such programs. In this presentation we will set out a teacher training data literacy model for informing instructional knowledge and practice. The model will cover definitions and regulatory expectations for data literacy, where to find available data, examples of how several School Districts (SDs) are using data to support student learning, and hands-on work with real student data.

# Navigating Change in Chile's Higher Education Access System: Examining Standardized Testing and Governance Amid Societal Unrest

D. Jimenez[1], M.L. Varas[1]

[1]Universidad de Chile, Chile

The objective of this presentation is to examine and discuss the changes implemented in Chile's highly centralized and test-based higher education admissions system. Although standardized testing has a long tradition in Chile dating back to the 1960s, recent modifications to the assessments and system have been introduced amidst growing national and international scrutiny and skepticism towards standardized testing. We analyze these changes and their effects from the perspective of the institution responsible for developing the assessments.

Our presentation specifically explores the pressures and expectations placed on the admissions system and tests by various social groups. We trace the nearly 25-year history of institutional actors' attempts to address these demands, the outcomes and limitations of their responses, and how these efforts have shaped the current changes to the tests.

We conclude by discussing potential strategies for responding to the evolving challenges and societal demands for justice and equity required of institutions responsible for high-stakes, large-scale standardized assessments.

# Assessment Cultures II

9:00 - 9:30

## Exploring the longitudinal development of assessment practitioners via their participation in assessment professional development.

H. Williams[1], S. Child[1]

[1]Cambridge University Press and Assessment, United Kingdom

A crucial element of successful assessment reform is ensuring an increased rate of assessment literacy among education professionals, particularly for teachers and school leaders. Assessment literacy can be understood as knowledge of the principles and practice of assessment, including techniques, purposes and processes. Professional development for education professionals within the context of assessment reform and innovation is vital, yet often overlooked in research. This study examined the longitudinal impact of two assessment-related professional development courses on participants' professional lives, discussing in particular their performance, confidence, career development, and engagement with assessment as a whole. This longitudinal mixed methods study took data snapshots from three time periods: before participants undertook the courses, immediately after, and 12-24 months following course completion. We analysed questionnaires and feedback and undertook semi-structured interviews with participants from a broad pool of professions including teachers, school leaders, exams officers and assessment managers in the UK context and internationally. Participants perceived that completing professional development directly led to career progression and that this was attributed to socio-cultural interpretations of assessment literacy including self-efficacy, confidence and motivation. These findings are discussed in relation to how assessment professional development can be embedded most effectively to support educational reform.

# The impact of curriculum and assessment reform on practices in a high-stakes examination system: A Maltese case study

D. Pirotta[1], O. Vassallo[1]

[1]UNIVERSITY OF MALTA, Malta

One of the aims of the Learning Outcomes Framework in Malta sought to reform a national assessment policy through a change in assessment culture with a move away from an exam oriented approach. A change in assessment practices is considered to be key in order to generate a domino effect on teaching and learning practices in schools. The paper presents the journey that led to a significant overhaul to MATSEC's Secondary Education Certificate [16+] assessment design, mainly by developing learner centred syllabi which give importance to different forms of assessment and grant more flexibility to teachers. We will report on three main challenges that MATSEC contended with in the past six years: (i) the design of an assessment model and how it evolved from moderated coursework assessment to school-based assessment; (ii) achieving a balance between a strong exam culture with a colonial history and a more learner centred orientation; (iii) the alignment of assessment principles and practices with teacher judgement. In the end, the result is a compromise between validity, reliability and manageability.

# What can constructs of high stakes exams tell us about assessment cultures? The case of the new Language arts exam in Norway.

G.B.U. Skar[1], A.J. Aasen[1]

[1]NTNU, Norway

In 2020 the Norwegian Writing Centre was commissioned by the Norwegian Directorate for Education and Training to develop av new high-stakes language arts exam. The Writing Centre was also commissioned to, for the first time in the exam's century long history pilot both tasks and criteria. The backdrop was a curriculum reform and a white paper that evaluated the exam system and concluded with severe threats to validity.
This paper will detail the process of designing a new exam and the ensuing results. The process of designing a new exam has been characterized by a balance between demands of psychometric qualities and, as it were, test user perceptions. During the developmental stag the NDET has published "suggestions" to new tasks and scoring models with the intent of receiving input from test users (e.g., teachers, students). Teacher associations as the Association of Language Arts Teachers in Norway, and teacher unions has been granted strong input in the decision-making process of the new exam and these inputs have to a non-trivial extent challenged test drafts designed to increase psychometric qualities of the exam. After three year of development, the NDET has decided to retain the 2019-version of the language arts exam.

## Technical, Vocational and Applied Assessments II

9:00 - 9:30

### Strategies to allow multiple voices to be heard in assessment reform: Engagement strategies and their findings from Qualifications Wales' review of qualifications in the Travel, Tourism, Hospitality and Catering Sector.

L. Mitchell[1]

[1]Qualifications Wales, United Kingdom

In March 2023 Qualifications Wales published its report on its review of qualifications, and the qualification system in the Travel, Tourism, Hospitality and Catering sector. This was the fifth of a series of reviews carried out by Qualifications Wales, the regulator of qualifications, other than degrees, in Wales. The purpose of each of these reviews was to identify the extent to which, in that sector, the needs of employers and learners were being met – with a focus on learners aged 14-19 in full-time education and on those following apprenticeships. A key focus for each review has been to identify whether the assessment arrangements are effective. Using a mixed methodology of semi-structured interviews, focused discussion groups and an online questionnaire, Qualifications Wales gathered the views of a range of interested parties such as learning providers, employers, sector experts and learners. As the review was conducted during the second year of the COVID-19 pandemic, all interviews were held online. While some of the focused discussion groups for learners were also undertaken online, others were conducted face-to-face. This paper outlines the methodology used and the findings of the review, which will be used to inform a programme of qualification reform in Wales.

# Lessons learned from working with partners to introduce vocational qualifications with significant project-based assessment in sub-Saharan Africa.

P. Ashton[1]

[1]Cambridge University Press and Assessment, United Kingdom

Education reforms by the Governments of Botswana and Eswatini introduced new suites of examinations focussing on technical and vocational skills as well as academic competencies (education with production). They were to be of equal value to the existing national standards (IGCSE). The reforms aimed to produce learners who were competent in both academic and vocational skills, with appropriate competencies for self-employment, employment, and further training. In addition, the new syllabuses should foster entrepreneurial skills and the integration of interpersonal and application skills.
This paper discusses the lessons learned so far from working the exam boards to develop and implement an assessment model, assessment tools, and standards.
Challenges included finding enough teachers with pre-requisite subject knowledge, allocating teachers to schools, providing training, encouraging take up by schools, securing physical and teaching resources. Schools had to decide which of the existing syllabuses to drop, find and equip space, and enrol students. Such changes risked conflict with teaching unions.
Assessment models were developed which realised the governmental vision and mitigated against the challenges.
This was a challenging but worthwhile reform, requiring significant resources and cooperation to enact. Assessments need to be robust and be appropriate to the prevailing capacity, risks, and limitations.

# Feasibility of using z-score mark estimations for new and existing Technical Qualifications

Z. Rahman[1]

[1]City & Guilds, United Kingdom

The estimation of marks continues to play an important role for learners, centres, awarding organisations and other stakeholders, in enabling the continuation or completion of study without causing unnecessary delays where possible. A programme of research was undertaken to investigate the applicability and accuracy of the z-score method for vocational qualifications. The first phase involved an evaluation of marks being estimated using the z-score, percentile and proportional methods; the second phase evaluated the quality of z-score estimates over time and using a databank of marks from multiple series; and the third phase reviewed the impact of using estimated assessment marks on the overall qualification results.

This paper presents a summary of findings from this research and demonstrates that z-scores can generally be used to estimate marks for Technical Qualifications. However, it also presents evidence from an evaluation of estimated marks for T-levels qualifications to illustrate how the z-score method may not always be suitable for new qualifications. This can be a particular challenge, especially given that technical and vocational qualifications have long been the object of major government initiatives and reforms, which are sometimes introduced at pace and could at times lack the opportunity for comprehensive technical reviews.

## Formative Assessment II

9:00 - 9:30

### Moving towards a new assessment culture in Malta: the implementation of School-Based Assessment in the French as a Foreign Language classroom.

R. Bonello[1]

[1]University of Malta, Malta

Reforms in the educational sector within the Maltese context are ongoing (Eurydice, 2023). The launch of the syllabi based on the Learning Outcomes Framework (LOF) which commenced in 2018 gave rise to new assessment practices.
The present study is conducted in Maltese Secondary Schools as part of a larger project titled “Assessment of French as a Foreign Language at Secondary Level in Malta”. Its main aim is to document the period of transition in order to corroborate or refute teacher empowerment with regard to assessment advocated by the Learning Outcomes Approach. Data collection took place in two state schools and two church schools. Year 9 and Year 10 lessons spanning over a unit were recorded and transcribed. Excerpts of lessons are being analysed and examined in the light of assessment for learning practices (Wiliam & Jones, 2007) and pedagogical assessment systems (Macaro et al, 2016; Pasquini, 2021) devised to enhance the power of assessment in the modern foreign language classroom (Pace, 2018). The research paper hence attempts to find out whether School Based Assessment is serving the purpose of formative assessment in the FFL classrooms and the roles that teachers and learners are playing in the learning and assessment processes.

# Ringing the inner voice: Students' Experiences of Teacher-Feedback using Narrative Frames.

F. Passeport[1]

[1]University of Dundee, United Kingdom

In this session, we will examine the use of a narrative inquiry method called Narrative Frames within the context of a study on undergraduate students' experience of teacher-generated feedback.
Experiences are hard to observe and limited to make sense of with a questionnaire approach. In addition, researchers can sometimes feel limited in their conclusions when following a purely qualitative approach (such as interviews) with a small sample. Therefore, the use of Narrative Frames may provide an opportunity to gather considerable data without compromising its quality.
Discovering this method may provide researchers with new insights into how to elicit hidden evidence, such as unspoken and untold stories, that enrich our understanding of the invisible processes and experiences of students or teachers.
Supported by the application of the method and empirical evidence, the session will primarily focus on designing research with Narrative Frames, through the examination of the study's exemplar, and will engage the audience in interactive exercises to unpack the different elements and principles of this approach.

## Fairness & Social Justice II

9:00 - 9:30

# Measurement invariance across educational systems in the First and Second International Science studies

Y. Sosa Paredes[1], B. Andersson[1]

[1]Centre for Educational Measurement (CEMO), Norway

This work investigates measurement invariance across educational systems in the First and Second International Science studies (FISS and SISS, respectively) conducted by the International Association for Evaluation of Educational Achievement in 1973-74 and 1983-84. In previous studies, researchers have examined how well the tests measure science achievement, their psychometric properties, and how the mean science achievement changes across time. However, it is unclear if the tests are consistently measuring the same construct across different educational systems and over time. This study applied classical item analysis, multiple group item response theory modeling, and the alignment method to address this question. Our findings showed that some items in the assessments had negative point-biserial correlations in at least one educational system. We also found that a unidimensional two-parameter logistic model fitted well in most educational systems in both studies. While we established configural invariance of the tests across a majority of the educational systems, we detected a lack of scalar and partial invariance across these educational systems in both assessments. To evaluate the robustness of our findings we considered approximate measurement invariance via the alignment method, which did not identify any items that had invariant parameters across all educational systems.

# Education Reform: Improving Educational Prospects for Girls in India

I. Bhaduri[1], D.P. Saklani[1], P. van Rijn[2], H. Por[3], J. Bertling[3], K. Ghosh[1]

[1]National Council of Educational Research and Training, India
[2]ETS Global, Netherlands
[3]Educational Testing Service, USA

Assessment reforms are often motivated by the changing needs of society. In India, the education policies highlight the need for a more equitable learning environment between boys and girls. The Beti Bachao, Beti Padhao (BBBP, translated as "Save the Daughter, Educate the Daughter") initiative was launched in 2015 to address the declining ratio of girls to boys in India. Jointly run by the Ministry of Women and Child Development, the Ministry of Health and Family Welfare and the Ministry of Education, the initiative also sought to increase girls' participation in school.

This presentation provides an overview of the BBBP initiative and its impact on the learning achievement of boys and girls as measured by the National Achievement Survey (NAS) administered in 2021 to a representative sample of students across all states in India. In particular, we will discuss the NAS learning achievement levels in Punjab and Haryana. These states are of special interest because the girls in these states outperformed the boys. For each state, we will focus on the learning achievement of three districts where the girls outperformed the boys by a large margin and three school districts where the girls outperformed the boys marginally.

# It’s a question of style: Understanding learner interactions and preferences with text styling in Onscreen Assessments

E. Barrow[1], E. Crampton[1], I. Custodio[1]

[1]Pearson, United Kingdom

This presentation explores one aspect of a multi-phase programme of research aiming to improve the accessibility, inclusivity, and usability of high-stakes onscreen assessments. Specifically, this research focuses on the optimisation text styling and formatting. Reviews of our current onscreen assessment content and platform, as well as technical accessibility and usability standards and guidelines, identified existing inclusivity issues and formed an initial set of recommendations to improve the accessibility and user experience of our onscreen assessments.

User trials were conducted with students to seek feedback on the recommended improvements to onscreen text styling and formatting. Students were asked to complete two onscreen English tests, one where these improvements had been implemented and one where they had not. Feedback indicated an overwhelming preference for the test containing the improvements, with evidence suggesting a positive correlation between the implemented recommendations and improved user experience. We suggest that these improvements can benefit the overall test-taking experience for all learners while also inadvertently catering to the needs of many students who may have undiagnosed SEND conditions. Moreover, by prioritising a user-centred approach to the development of future onscreen assessment, we can also ensure that student voice underpins the decisions we make.

# Discussion Group 1

11:00 - 12:00

## What are the possible consequences of giving teacher assessment a large(r) part to play in school student certification?

I. Nisbet[1], S. Shaw[1, 2], L. Wiseman[3]

[1]University of Cambridge, United Kingdom
[2]University College London, United Kingdom
[3]Independent consultant, United Kingdom

The focus of the Discussion Group is the ‘intentions’ to ‘enactment’ stage of assessment reform journeys. It is intended to provide an opportunity for participants to explore cultural, contextual and individual factors at play, among diverse stakeholder groups, that could have critical consequences for the enactment of any significant educational assessment reform in any part of the world. A number of countries have revised national, end-of-school assessment systems to give a larger role to teacher assessment. The successful implementation of such reforms must recognise and negotiate the cultural positions, concerns and expectations of all stakeholders. This session will provide an engaging opportunity for participants to explore those positions for different stakeholder groups and the impacts upon them of a realistic scenario in which a national decision has been made to increase the role played by teacher assessment in the summative assessment of students at the end of compulsory schooling. Organised in different groups, each with the role of a single category of stakeholder, the groups will be provided with the same scenario. To prompt discussion, each group will also be provided with relevant stakeholder questions. A plenary will provide opportunities for sharing key issues and benefits identified by respective stakeholders.

# Discussion Group 2

11:00 - 12:00

## Assessment Reform Journeys: Post Graduate Students and Early Researchers Across Borders

J. Leonardsen[1], D. Normann[2], S. Manassian[3], S. Vassiliou[4], G. Cipriano[5]

[1]NTNU, Norway
[2]norwegian university of science and technology (NTNU), Norway
[3]PSI Services UK Ltd, United Kingdom
[4]University of Cyprus, Cyprus
[5]ISCTE-IUL, Portugal

The aim of the Post Graduate Student and Early Researcher Network is to create a supportive and caring environment where the members can find a community of people experiencing similar challenges. This need underpins the discussion session that the steering group would like to propose. The Network would like to invite all current members and any interested participants at the conference to join us at a discussion session where we will explore the conference theme Assessment reform journeys: intentions, enactment and evaluation in relation to our experiences as students and as assessment researchers.

Our network includes post graduate students and early researchers from all over Europe. We see that the education systems in our home countries have both similarities and differences based on political and social context. Educational policy reforms are usually instigated by the changing educational needs. However, the need for changes in educational systems varies from country to country. This discussion session will allow us to explore assessment reforms in different European countries, for example Portugal, Cyprus, Scotland and Norway, and discuss opportunities and challenges across borders. The key areas for discussion are: motivation for assessment reforms, tensions in assessment reforms and the impact of assessment reforms.

## Discussion Group 3

11:00 - 12:00

# Organizing Effective Thesis Calibrations with A Typology of Calibration Methods

Y.P. Hsiao[1], A. Kramer[1], G. van de Watering[2]

[1]Tilburg University, Netherlands
[2]Eindhoven University of Technology, Netherlands

The graduation thesis is a critical research undertaking that marks the end of a student's program of study. To ensure thesis assessment quality, conducting calibrations (also called moderations) is an essential approach in which a group of examiners discusses their interpretations of an established assessment instrument, such as a rubric. However, there is a lack of research on how to calibrate assessment practices effectively in the context of a graduation thesis. Previous studies have focused primarily on examiners' consistency, while the broader context in which thesis assessments are designed has been overlooked (e.g., alignment between thesis learning goals and program learning outcomes and curriculum, examiners' expertise). Effective dialogue during calibration sessions is also crucial to promoting a shared understanding of assessment practices and fostering a sense of community among examiners. To address these issues, the authors propose a typology of thesis calibration methods that considers goals, facilitation, participants, context, interaction, and improvement. The discussion will start with participants' past experiences in calibration sessions, followed by exploring the typology's five dimensions and how they contribute to effective calibration and thesis assessment quality. Finally, participants will brainstorm how this typology can be used to cultivate assessment culture at the thesis level.

# Discussion Group 4

11:00 - 12:00

## What do assessment professionals think should be included in a code of ethics for using process data in educational assessment?

D. Murchan[1], F. Siddiq[2]

[1]Trinity College Dublin, Ireland
[2]University of South-Eastern Norway, Norway

The potential of process data derived from educational assessment has gained much attention during recent years. However, possible ethical concerns related to the use of such data appear to lag behind the technical developments. Therefore, in this discussion we will initiate a debate on this topic amongst the assessment community. Discussion will focus on participants' views on the relevance of a recently developed (work in progress) code of ethics for using process data, other novel data sources and associated methodologies for exploiting such data in the field of educational digital assessment.
This proposed discussion aligns with the conference theme, problematising the gap between intentions and enactment in relation to one emerging practice in assessment—use of process data—and links to the questions regarding equity, fairness, trust, and accountability in the Call. We propose a discussion in which we address questions such as: Given a set of possible ethics standards, how does each dimension suit your context; Are there elements of ethics that apply differentially across professional contexts (e.g., researchers, test developers). We anticipate that the discussion will raise awareness among participants and broaden the work in this important area on the ethical use of process data in educational assessment.

# Discussion Group 5

11:00 - 12:00

## Beyond the hype – understanding the limits and potential of AI in education

C. Aloisi[1], I. Casebourne[2], R. Hamer[3], C. Tupling[4]

[1]AQA, United Kingdom
[2]DEFI The Bridges Hughes Hall, Cambridge University, United Kingdom
[3]International Baccalaureate, Netherlands
[4]AQA Education, United Kingdom

The aim of this eAssessment SIG discussion group is to move beyond the ‘crisis’ narrative and focus on the shortcoming as well as the promises of real-world AI tools and systems. The panel members will introduce four perspectives on specific current limitations of AI and LLM (drawing from their poster submissions). They aim to open the discussion on how these inherent limitations may affect the deployment of AI in education and educational assessment and what can be done to overcome them. Such discussions will prove useful – and perhaps even essential – to the efforts to bring different stakeholders together and develop a common language. A shared language will support an effective, responsible and inclusive introduction and embedding of AI in high-stakes education.

The panel will pose initial questions regarding the impact of AI in education and society. More than half of the session is to be devoted to small group discussion. Possible initial opening questions include dimensions of AI literacy, AI's ability to represent diversity, creativity and innovation, valid and trusted ways to use AI now, what is needed to make them more useful for inclusive education, and some misunderstandings between the different stakeholders involved with AI in education.

# Discussion Group 6

11:00 - 12:00

## Removing unnecessary barriers: practical considerations for designing accessible digital assessments.

D. McVeigh[1], E. Barrow[1], I. Custodio[1], E. Crampton[1]

[1]Pearson Education, United Kingdom

As we transition towards digital assessments for national examinations, it is crucial that exam boards and governmental bodies consider how we can make digital assessments as valid, reliable and fair as possible. A central facet of doing so is in the removal of barriers to access that may currently exist within digital assessments. For this discussion group, we will explore how we can better consider accessibility in the context of digital assessment by asking the following questions:

How can digital assessment benefit students with accessibility needs?
How can we improve the design of our assessments and structure of our items to improve their accessibility?
How can accessible digital assessment design improve equity of access?

These questions will underpin our discussion, allowing participants to contemplate accessibility in a more holistic way, from the impact on learners to the decisions we should be making in the design and development of digital assessments. In forming the discussion in this way, we hope to build participants' awareness of the different aspects of accessibility that need to be considered. For instance, emphasising the importance of both content accessibility (test design and content) as well as native accessibility (web and platform capabilities).

## Discussion Group 7

11:00 - 12:00

# Digital Formative Assessment: dialogue, implications, and ethics in the context of a European policy experimentation project

J. Elwood[1], K. Livingston[2]

[1]Queen's University Belfast, United Kingdom
[2]University of Glasgow, United Kingdom

Assess@Learning (2019-2023) was a European policy experimentation that considered the implementation of digital formative assessment (DFA) in schools. The research approach included Country Dialogue Labs and Student Dialogue Labs, bringing together multiple stakeholders in dialogue to gather data on the topic of DFA.

This discussion group will explore:
(i) The dialogue lab approach in policy development - How might the Dialogue Lab approach be used in different contexts to enable stakeholders, including students, to engage in digital assessment policy development?

(ii) The implications for students of digital formative assessment - What are practical ways we could work with students to enable their greater involvement in developing digital assessment policies?

(iii) The ethical implications of digital assessment practices - What are the ethical implications of greater use of digital assessment - especially the use, storage and sharing of digital assessment data - for policy makers, test development agencies, and schools?

Stages of the research have been presented at previous AEA-Europe annual conferences. We wish to conclude this journey and discuss key final considerations and implications with participants at the 2023 conference with a desirable outcome being a greater understanding of the implementation of digital assessment in the various European contexts represented.

## Discussion Group 8

11:00 - 12:00

# Mathematics assessments for the future – taking into account the communicative and collective aspect of mathematical reasoning

G.A. Nortvedt[1], L. Sumpter[1]

[1]University of Oslo, Norway

An overarching aim of mathematical activity is to “pose and answer questions in or by means of mathematics” (Niss & Højgaard, 2019), illustrating the social aspect of mathematical competence and that competencies are displayed in interaction with other individuals. Most assessment techniques build on the idea that assessment should be individual in order to be as objective as possible, to get a ‘fair’ description of the individuals competence. One example of a mathematical competency is mathematical reasoning, the ability to justify choices and conclusions by arguments. Researchers stress the social component in mathematical reasoning, meaning that justifications of choices and conclusions are in relation to a specific context where other actors are present. As a social activity, it aims for meaning making. We argue that such meaning making cannot be assessed in a traditional, individual exam. Moreover, reasoning is merely one of several competencies that comprise collective and social aspect. In conclusion, the current situation with the enhanced attention to mathematical competency calls for new and different ways to assess competencies such as mathematical reasoning. The aim of this discussion group is to discuss affordances of individual assessments vs collective assessments to assessing mathematical competence.

# Assessment Cultures III

14:00 - 14:30

## A Critical Reflection on the Implementation of Programmatic Assessment

L. Duchi[1], F. Passeport[1, 2]

[1]Erasmus University Rotterdam, Netherlands
[2]University of Dundee, United Kingdom

Programmatic Assessment is a model that is gaining popularity in higher education as a solution to the traditional formative-summative assessment approach. With Programmatic Assessment, the focus is on the student's growth journey, represented by longitudinal data points from low-stake assessment moments throughout a course. Teachers take the role of coaches who support students in setting goals, and in progressing through them with reflective conversations and feedback moments. Eventually, the data points are gathered into a portfolio that serves as the tool to tell a student's unique learning story and provides the basis for the independent committee to make a high-stake decision (pass/fail or a grade). In this paper, we will expose our critical reflections on implementing Programmatic Assessment in a Minor course on 'The Future of Technology' in 2020 and 2021 by sharing the concrete challenges and opportunities we experienced and making recommendations for practice.

# The strange non-death of SIMCE tests: The multiple survival strategies of a controversial and contested market mechanism

E. de Padua[1, 2], M. Rosenzvaig[1]

[1]University of Cambridge, United Kingdom
[2]SUMMA, Chile

SIMCE is the national system for evaluating the quality of education in Chile, which has been in operation for over 35 years, one of the oldest large-scale evaluation systems in Latin America. Despite its long history, it has remained a controversial national educational policy. This paper aims to examine the evolution of SIMCE, from its creation during Pinochet's dictatorship in the 1980s to the present day, and explore how its most controversial features have persisted despite social movements and educational reforms.
To achieve this, the study examines official documents from the Ministry of Education, media publications, and interviews with key stakeholders involved in the design of national policies. The analysis reveals that SIMCE's roots are embedded in an educational system where quality and access are defined by a marketized approach. The paper concludes that SIMCE has evolved in its visibility, intensity, and uses over the years, but its controversial nature remains. It highlights the importance of critically examining the educational policies that shape national education systems and underscores the need for alternative approaches that prioritize equitable access and quality education.

# How do students experience their teacher’s didactical practice of formative assessment?

E.W. Hartberg[1], K.K. Bjerke[1], T.S. Wille[2], T.E. Wiig[2]

[1]Inland Norway University, Norway
[2]Oslo Municipality, Norway

The Norwegian enactment of assessment was revised in 2020, and it clearly shows that the assessment shall contribute to the students «desire to learn», and that formative assessment should be seen as an integrated part of the students learning process.
The Inland Norway University has been in a professional partnership with a selection of secondary schools in Oslo municipality for more than three years, working with improvements of their formative assessment practises. Each school has included a group of students in their work and have systematically collected their views and perspectives on assessment and learning.
In our qualitative study, we seek to describe how schools work with formative assessment in line with the regulatory changes, and how the students experience their teachers’ practices. We have raised three research questions that we will address in this paper:
1. How have the schools been working to improve their formative assessment practices within their professional learning communities?
2. How do the students experience their teacher’s didactical practice in formative assessment?
3. How does the summative assessment at the end of the course affect students and teachers understanding of formative assessment during the course?

## Psychometrics & Test Develpoment III

14:00 - 14:30

### Attitudes to the importance of empathy in police work: the student perspective

T. Stenlund[1], C. Wikstrom[2], M. Inzunza[1]

[1]Umeå University, Sweden
[2]Umea University, Sweden

In many educational programmes preparing students for professions where human interaction is central, the development of empathic ability is of importance. The aim of this study was to explore attitudes towards the importance of empathy in police work among police students. The intention was also to investigate if attitudes would change over time during education, and whether gender differences could be observed. A modified version of the Jefferson Scale of Physician Empathy was administered to 355 Swedish police students in a mixed method design, including both between- and within-groups comparisons. Attitudes toward the importance of empathy were measured before and after a practical internship. The result showed that the students in general found empathic ability important, but also that they did not change their attitudes over time. The results were conflicting when it came to gender differences. There was no significant result for the between-group sample, but female students reported significantly higher compared to male students in the between-within group. This research is valuable in the discussion on how “soft skills” such as empathic ability can be assessed and developed in students in preparation for professional practice, and can also be useful in educational evaluation and planning.

# Measuring educational constructs qualitatively

A. Scharaschkin[1, 2]

[1]AQA, United Kingdom
[2]University of Oxford, United Kingdom

Louis Thurstone claimed that 'when the idea of measurement is applied to scholastic achievement, ... it is necessary to force the qualitative variations [in learners' performances] into a quantitative linear scale of some kind'. The psychometric approach to educational measurement is the logical development of this view. It depends on a conception of measurement that equates measuring a phenomenon with finding the value of a quantity, i.e., a location on the real number line.

Bas van Fraassen expands the idea of measurement to conceive of it more broadly as location in a logical space. This talk will argue that van Fraassen's view is more consonant with much educational assessment practice in England. Here, assessment constructs are thought of as qualitative amalgams of subject content and assessment criteria, with respect to which (fuzzy) concepts of levels of attainment are defined. To measure a learner is not to estimate how much ability she has, but rather to locate her in a partially-ordered logical space, based on the extent to which she has demonstrated attributes of attainment associated with a given level.

A mathematical approach to this problem will be outlined, and analogues with quantitative latent variable methods will be drawn out.

# Investigating the Robustness of DIF-Analysis in Standardized Testing: A Longitudinal Study of Experimental Items

I. Laukaityte[1], P. Lyrén[2], C. Wikstrom[3]

[1]Umeå university, Sweden
[2]Umeå University, Sweden
[3]Umea University, Sweden

In all tests, and especially tests with high stakes, the test items must be relevant to the construct or domain being measured but also fair, in the sense of being invariant across groups of test takers. Differential item functioning (DIF) analysis is used in tests to identify systematic differences in item performance among test takers. Various methods can be used to investigate DIF, each with their own strengths and weaknesses, which also can make it challenging to interpret findings. In this study, we aimed to investigate the robustness of DIF-analysis by comparing the outcome of different methods over time. We focused on experimental items from a standardized verbal ability test used in the admission to higher education in Sweden, that were administered for seven years across 13 test administrations. Almost 50 % of the studied items showed DIF at some point, randomly distributed over time. Additionally, the results indicated that items that measured reading comprehension and higher-order skills were more consistent over time, whereas vocabulary items exhibited more variability. We also examined selection effects and sample size by creating comparable groups based on test-taker characteristics, without a changed outcome. We will discuss the findings and possible explanations.

## Assessment that is reactive to unforeseen circumstances (e.g. Covid 19) IV

14:00 - 14:30

### Evaluation of the 2022 Approach to the Assessment of Graded National Courses in Scotland: Learner and Practitioner Experiences

S. Allan[1], S. Hill[2]

[1]Scottish Qualifications Authority, United Kingdom
[2]SQA, United Kingdom

Evaluation of the 2022 Approach to the Assessment of Graded National Courses: Learner and Practitioner Experiences seeks to understand learners and practitioners' experiences of the assessment of National Qualifications in Scotland in 2022, and to contribute to the ongoing debate on how best to assess young people.

It looks at all parts of the assessment process that had not yet returned to the approaches taken prior to the pandemic, including modifications to assessment, revision support and the appeals process, as well as comparing both learners' and practitioners' views of the systems used to assess graded National Courses in 2021 and 2022.

In this research, a mixed-method approach was adopted, consisting of an online survey distributed to all schools and further education colleges in Scotland offering National Qualifications, followed by a range of semi-structured qualitative interviews with learners and practitioners. This approach allowed for further, more detailed exploration of points of interest from the survey results, as well as ensuring that different questions were answered fully and effectively. Over 2,000 learners and 1,000 practitioners completed the survey, and 40 depth interviews were carried out.

# An evaluation of post-16 maths qualifications in England: gathering evidence on standard setting and maintenance to inform policy change

N. Zanini[1], M. Wayman[1], T. Bruckbauer[1]

[1]Ofqual, United Kingdom

Promoting education and training in STEM disciplines has become a pressing policy priority for many governments. In England, the need to increase the provision of maths post-16 as the educational long-term response to boost economic growth and raise standards has been a major feature in the education policy debate of late.
The aim of this research project is to gather further evidence on post-16 maths provision in England, through an in-depth analysis of qualifications taken in upper secondary schools and the characteristics of students taking them. We also provide robust quantitative evidence on the setting and maintenance of maths qualifications standards to evaluate the potential impact on learners taking different maths qualifications or at different points in time.
Exploiting historical micro-data on upper secondary school students in England, we first use multilevel modelling to provide insights into the factors affecting students taking alternative post-16 maths qualifications. We then rely on a statistical definition of comparability, implemented through the use of propensity-score weighting, to evaluate the maintenance of qualification standards.
Our findings will be used alongside existing evidence to discuss the current provision of post-16 maths in England and inform policy making.

# Fairness & Social Justice III

14:00 - 14:30

## Learner preferences and inclusive assessment

A. Harrison[1], V. Rotaru[1]

[1]Qualifications Wales, United Kingdom

Inclusive assessment presupposes that the design and implementation of assessment consider the needs and interests of all learners. The design of assessment is often regarded to cause unfairness, particularly in a context where qualifications are used for high stakes purposes. When considering whether qualifications are inclusive, an important perspective to capture is the views of learners. We conducted interviews with learners to investigate the impact of various assessment aspects on their ability to demonstrate what they know and can do. The research showed that the impact of factors that make assessment engaging, hence inclusive, is not the same for all. It also highlights that preferred assessment approaches are not without drawbacks and that designing qualifications and assessments must consider the trade-offs. This presentation will describe selected findings from the study and will map them to the findings from a desk-based review of inclusivity within our qualifications system, as well as wider research in this area. We will consider the extent to which learner preferences are reflected within our own system and how these preferences could be considered when designing an inclusive assessment system.

# Amplifying Small Voices in the Age of Big Data: A Qualitative Study of Vision-impaired Students on the Use of Modified Exam Papers and Assistive Technologies

L. Liu[1], K. Mason[1], B. Redmond[1], H. Dalton[2]

[1]Pearson, United Kingdom
[2]Pearson Education, United Kingdom

This paper aims to unfold England vision-impaired (VI) students' holistic exam-taking experience and understand the challenges they met during high-stake exams at the General Certificate of Secondary Education (GCSE) level. Using an exploratory qualitative approach, this study firstly interviewed seven year 9-12 students from two schools in England - one mainstream and one specialist, and then augmented by brief interviews with teaching professionals working with students with VI.

The qualitative data shows questions with graphs, including maps, diagrams, and images, are most challenging for students with VI during the exams. We found assistive technologies (ATs), including physical mathematical models (visual & tactile) and text-to-speech (auditory), could significantly improve VI students' exam performance. The former increases stereopsis, enabling students to observe 3D shapes in a comprehensive way without eyestrain, while the latter reduces students' reliance on human readers, boosting their autonomy and confidence in taking exams. The result also indicates there is a discrepancy between school modified exam materials and modified exam papers (MEPs). Ensuring students have enough practice with suitable templates will empower them during high-stake exams. The paper contributes to the field by unfolding a small group of students' individual holistic experience of using MEPs and ATs.

## Higher Education & Assessment II

14:00 - 14:30

# The process of developing a personality questionnaire for screening candidates for higher education

A. Moshinsky[1], D. Gilon[1], D. Ziegler[1], G. Soffer[1], E. Ben Barak[1], M. Fronton[1]

[1]NITE, Israel

Over the past two decades there has been increasing emphasis on evaluation of non-cognitive traits when selecting candidates for certain professional study programs. Candidates with outstanding cognitive abilities are not necessarily suited to a particular profession and do not always become the best professionals. Advanced technology has enabled digital and online administration of personality questionnaires, allowing for easier and more efficient data collection and analysis. This has also led to the development of adaptive testing methods that can tailor the questionnaire to the individual respondent. Statistical techniques such as item response theory and network analysis have enabled advanced scoring and analysis.
These developments have made it possible to administer and score personality questionnaires easily, and relatively quickly and cheaply. Given these benefits, personality questionnaires have become a sought-after assessment tool for organizations seeking to evaluate candidates' non-cognitive skills.
However, personality questionnaires also have major drawbacks. These include:
1. Self-Reporting Bias and Response Bias
2. Clinical Utility and Ethical Concerns
3. Cultural Bias and Gender Bias
4. Limited Reliability and Validity
This presentation describes various ways in which these drawbacks can be managed. Data from a validity study are also presented.

# Rethinking a Higher Education Online Curriculum using Backward Design (Learning Outcomes-Assessment-Activities)

M. Barba[1], R. Estrada[1], L. Garelli[2], C. Martinez[2], P. Sánchez[2]

[1]Anahuac Online, Mexico
[2]Anahuac University, Mexico

This paper explores the work carried out in the design and implementation of a new curricular model for online graduate programmes. The intention was to assess the graduation profile and learning outcomes. Firstly, the context and background that initiated this project are explained, as well as the need to develop a new curricular model that would allow the assessment of the graduation profile of over 5,000 graduate students of online programmes. Secondly, some of the theoretical inputs that served as a reference for the development of the curricular model are presented, considering the contributions made by the Minerva Project with respect to the structure of core competencies, and with a special emphasis on the anthropological worldview based on personalism.

Subsequently, the central structure of the proposal is described. This consists of a set of core competencies anchored to the described human dimensions, from which descriptors to be assessed are developed, allowing the attainment of professional competencies. These are then linked to the instructional design through activities that act as critical evidence of performance. The use of this methodology, defining learning outcomes, ways of assessment, and finally activities, is called Backward Design.

# Preparation for a high stakes language examination: practitioners' views from the field of shadow education

S. Tsiplakou[1], D. Tsagari[2]

[1]Open University of Cyprus, Cyprus
[2]Oslo Metropolitan University, Norway

The research presented in this paper examines through data from teaching practice and practitioner inquiry, the ways in which washback effects are produced and sustained with special reference to a high-stakes language test in Greece. We recorded and analysed the teaching practices of three experienced Greek language teachers offering private tuition to students in reparation for the Panhellenic examinations. We analysed 15 teaching hours, during which the students were trained in essay ("Composition") writing on specific topics or were given feedback on their work. The analysis of the teachers' discursive and teaching practices was supplemented by recordings and analyses of their own observations on their lessons (via a think-aloud protocol). This micro-level study sheds light on how test washback effects are generated and sustained by practitioners who may be fully aware of the detrimental effects of washback on literacy learning. Our data also show that in order to better understand the negative consequences of the washback effect, it is necessary to obtain sufficient data from different cultural and educational contexts, and to assess the import of the practices of all stakeholders in the process.

# E-Assessment IV

14:00 - 14:30

## Evaluation of the Cambridge International Digital Mock Exams Service

M. Kuvalja[1]

[1]Cambridge University Press & Assessment, United Kingdom

Cambridge International are now offering a Digital Mocks Service for some international GCSE and A level exams in preparation for live exams. As part of the iterative design approach, we run a range of research activities throughout the product's lifecycle. After the first launch in January 2023, we were keen to evaluate the delivery of the mocks service, and we used a particular methodological approach for this purpose. We collected user data from the test platform, user experience data, and validation evidence.

The findings that are produced as a result of the data collection and analysis can then be used to demonstrate the extent to which the service’s purpose was met, and to inform the continuous development and improvement of the service based on this evidence. The data collection plan required a collaborative approach with different parts of the business carrying out the proposed research activities as part of the iterative design process. This evaluation methodology aims to give product teams a structure for the evaluation so that the purpose of each data collection strand and the research questions are clear and linked to the objectives of the Digital Mocks Service.

# Piloting on-screen exams: a shift in mode and a shift in culture

J.M. Ryan[1]

[1]AQA, United Kingdom

In exploring the possibility of a shift from paper-based to computer-based exams, there is the need for us, as assessment practitioners, to investigate mode effects and methods of maintaining comparability of outcomes. From the perspective of students and teachers contemplating a transition from paper- to computer-based exams, however, such change represents not just a shift in assessment mode but a shift in assessment culture.

Since 2020, AQA has been piloting novel digital assessments in the summative space; in Spring 2023, a new series of pilots was delivered in GCSE Computer Science and in A-Level French. Participating students completed post-test surveys about their experience, and focus groups were conducted with a sample of students and teachers in order to gather feedback about their experience and their perspectives on the potential implementation of digital exams.

Findings from the two subjects clearly depict how the introduction of digital exams can easily converge with current assessment cultures, as in GCSE Computer Science, or can diverge, as in A-Level French.

# Explain yourself: Expanding validity evidence for automated scoring through explainability

S. Hughes[1, 2]

[1]Pearson, United Kingdom
[2]University of Cambridge, United Kingdom

Machine learning algorithms and artificial intelligence systems are increasingly given the power to make decisions that can have profound consequences across several domains of human life, including high-stakes assessment. Advances in machine learning technologies have made it possible to train machines to score complex assessment tasks, such as written essays, that had previously required human raters.

The methodologies for evaluating the validity of scores produced by automated scoring (AS) systems have not advanced at the same pace as AS technology itself. The primary criterion of AS system quality continues to be agreement with human raters. This presentation will outline the current practices in evaluating AS systems and explore recent developments in eXplainable Artificial Intelligence (XAI) research that have the potential to offer a new way to evaluate the construct validity of AS decisions. Findings from a pilot study on generating XAI-based explanations of essay scoring decisions will be presented to demonstrate the utility and limitations of these techniques. This presentation will be relevant to researchers, practitioners, and policymakers interested in the responsible use of AS systems for high-stakes assessments.

# Perspectives of End-users and the General Public on Assessment I

14:00 - 14:30

## Computer-based high-stakes assessments in England: Opportunities and Risks in the Eyes of Students and Parents

Y. El Masri[1]

[1]Ofqual, United Kingdom

The Office of Qualifications and Examinations Regulation (Ofqual) has conducted research examining students' and parents' perceptions of the opportunities and barriers of adopting onscreen assessment in high-stakes qualifications in England as well as their concerns over such a move. This research will inform regulatory approaches Ofqual adopts in the future.
A survey of mostly closed response questions was completed early in 2023 by a representative sample of 500 students, aged 15 to 18, who were preparing for high-stakes GCSEs and A levels and sitting their examinations in 2023 and 2024. A similar survey was also completed by a representative sample of 500 parents of such students.
The survey findings suggest that students are fairly comfortable and familiar with digital devices and most have access to them at home. Most students and parents reported being in favour of seeing some assessments move on screen. Students and parents appreciated the functionalities of onscreen assessment (e.g., spell-check, typing etc.) and were less interested in the potential flexibility in location and time this mode of assessment offers. Students' and parents' concerns were mainly around malpractice, data security and unequal access to technology.

# Review and reform of Essential Skills qualifications in Wales

G. Downey[1], P. Johnson[1]

[1]Qualifications Wales, United Kingdom

Essential Skills Wales qualifications – which includes communication, numeracy, digital and employability qualifications – have been around in their current form since 2015 but have existed in various guises (such as Key Skills and Basic Skills) for over 20 years. They are a core requirement within apprenticeships in Wales and are widely used in further education colleges as an alternative to GCSEs in English/Welsh and mathematics.

Essential Skills Wales qualifications have been the focus of reforms over the past 20 years and have featured different forms of assessment from teacher led assessment in the pre-2015 versions of the qualifications through to tightly controlled examinations and tasks in the current versions of the qualifications. However, the current versions are now eight years old and have been criticised by apprenticeship providers who have reported that the assessments can be unmanageable to deliver in the workplace.

Therefore, Qualifications Wales – the qualifications regulator in Wales – has undertaken a detailed programme of research to identify options for reform. This Open-Paper presentation will outline the findings from the first stage of the research and the options for the reform of Essential Skills Wales qualifications.

# Building a better qualification system: why the reform of construction and built environment qualifications in Wales demonstrates that effective reform must extend beyond qualification development

D. Seabrook[1]

[1]Qualifications Wales, United Kingdom

In this paper, the national qualifications regulator in Wales builds on research it undertook to review the effectiveness of its implementation of an ambitious suite of new qualifications in the construction and built environment sector following their first year of delivery. The paper acknowledges the scarcity of established models of effective implementation of vocational education reform as a driver for carrying out such research activity, and summarises the experiences of using the qualifications reported by learning providers during a programme of semi-structured interviews.

In the context of the learning providers' varied responses to the reforms, partly attributable to the localised nature of implementing any educational reform, the paper proposes that such mid-implementation research and review provides valuable opportunities for policymakers to adjust and refocus their approach to supporting qualification users. The paper concludes that while such research places resource demands on bodies who act to instigate and implement reforms, it offers opportunities to ensure the connectedness of live qualification delivery with the overarching reform aims, and affirms that in this field there may be no shortcuts to success.

# Other I

14:00 - 14:30

## New policies, old practices: The enactment of assessment reform in Maltese science classrooms.

R. Attard[1], D. Chetcuti[1]

[1]University of Malta, Malta

While the intended outcomes of assessment reform are clearly written in terms of policy, there is very little research about how these policies are enacted in practice and implemented by teachers at school level. This presentation will discuss the results of a small case-study that looked at the introduction of school-based assessment in science classrooms in Maltese state schools, as a replacement for summative half-yearly examinations. Using qualitative data from interviews with seven Maltese science educators, the presentation explores the decision-making process of the teachers as they navigated through the implementation of the changes in assessment practices.

The results of the study suggest that teachers responded to assessment reform by moving through different cycles of adaptability. They moved from trying to resist change by stating that the reform was simply a new name for old practices, to coping with change by trying to retain summative tests within the new framework of school-based assessment and eventually to an adaptability that empowered them to embrace new methods of assessment such as project-based learning, role-play and oral assessment. The presentation draws on these results to discuss the important role of teachers in the enactment of any assessment reform.

# Delivering Humanities Syllabi that address 21st century skills (problem-solving skills and communication skills): A Comparative Evaluation of the Ethics, Environmental Studies, Geography and Business Studies Syllabi and their assessment.

F. Zammit[1], R. Bartolo[1], A. Micallef[1]

[1]MATSEC, Malta

The presentation will explore the following questions: how do we identify and deliver 21st century skills, through formative and summative means, and how do we present subject specific syllabi able to attain these goals? What is assessable, and how can it be assessed to successfully address industry demands, while delivering a holistic educational programme ensuring the acquisition of transversal skills specifically through the Humanities subjects? The presentation will exemplify problem solving skills and communication skills as core skills within the 21st century skills discussion and proceed to discuss how problem-solving skills and communication skills are assessed in the Humanities subjects through a comparative examination of a selection of MATSEC Humanities syllabi. The aim of the presentation is to argue how the need for transversal skills can be achieved through the Humanities. Additionally, making the point that successfully including problem-solving skills and communication skills in subject specific syllabi can stimulate a positive impact on learners, and create learners who are well-versed with the requirements of a globalized world.

# Assessing the impact of vocational qualification reforms in England: Using large datasets to assess changes in outcomes for learners taking Applied General qualifications since 2016.

H. Dalton[1], S. Nastuta[1]

[1]Pearson Education, United Kingdom

This paper explores the extent to which Applied General qualifications (AGQs) have improved outcomes for learners in England since their introduction in 2016. It draws on data and analysis from a large-scale survey of teachers and learners spanning the period 2018 to 2022large national datasets and descriptive data of the changing AGQ cohort.

AGQs were created as the result of a wider policy reform in England aimed at simplifying post-16 pathways and were specifically designed to equip students with transferable knowledge and skills needed to continue their education through applied learning. Six years on from the first teaching of these qualifications, this paper looks at research and analysis done to understand the impact that this reform has had on outcomes for learners.

The outputs from the survey are contextualised by emerging analysis using large datasets which looks at outcomes for learners as they progress into the labour market and the extent to which there is evidence of impact from AGQs. Finally, the survey data is considered in the context of shifts in cohort characteristics over time and consider the extent to which this may affect outcomes and labour market impacts in the longer term.

# Symposium 1: Assessment for Social Justice

16:00 - 16:20

## Success and Growth for Every student

J. Zammit[1]

[1]Ministry For Education, Sports, Youth, Research and Innovation,, Malta

Formative assessment (FA) can bridge the learning gap by providing ongoing feedback to students throughout the learning process. Unlike summative assessments, which are typically used to evaluate student performance at the end of a unit or course, FA are used to monitor student learning and provide feedback that can be used to adjust instruction and support student progress. Teachers can identify areas where students may be struggling and adjust their teaching to better meet student needs. The teacher can provide additional support or re-teach the material in a different way to help the student understand. When students receive feedback that is aligned with learning goals and standards, they can use that information to identify areas where they need to focus their efforts and take steps to improve their learning.
Formative assessment can be a powerful tool to avoid using assessments that create tracking, where students are separated into different academic tracks based on their perceived ability, which can limit opportunities for certain students and perpetuate educational inequities. This approach allows for personalized instruction and avoids the need to separate students into different tracks. Formative assessment can also help to shift the focus from grades and scores to a deeper understanding of student learning. Instead of using assessments to rank and sort students, educators can use formative assessment to identify areas where students need additional support and to guide instructional decision-making. Overall, formative assessment helps by fostering a culture of continuous improvement in the classroom and to ensure that all students can succeed.

# Using assessment data to inform action in a state college literacy initiative.

D. Said Pace[1]

[1]Ministry For Education, Sports, Youth, Research and Innovation, Malta

Foreign language teaching and learning in the Maltese secondary school context is aimed as supporting learners reach different language proficiency levels while working together in the same classroom. This teaching and learning process requires a differentiated approach to pedagogy and assessment and seeks to provide a just and equitable teaching and learning environment for all learners. The discussion that follows in this regard is set against the backdrop of a process that addresses the need to further align FL curricula and assessment practices to principles of the Common European Framework of reference and its Companion Volume, initiated during the course of the LOF reform and the introduction of school-based assessment. Outcomes that emerged in the process of analysis of FL curricula and assessment conducted over the span of the reform and its implementation as part of the European Centre
for Modern Languages RELANG project and the Department of Foreign Languages within the Directorate for Learning and Assessment Programmes will be presented to inform the discussion. Inherent to the process described, and in the light of the introduction of school-based assessment, is the focus on the learners as active agents in the process of learning and assessment. Conclusions are drawn on the need to provide space for learner engagement through alternative forms of assessment processes within an action-oriented approach to language teaching and learning and the need to support learners and teachers to collate evidence of levels of language proficiency claimed through self and peer assessment.

# Opportunities for the development of foreign language proficiency through alternative assessment practices.

A. Micallef[1]

[1]Ministry For Education, Sports, Youth, Research and Innovation, Malta

The Malta National Literacy Agency (MNLA) carries out bi-annual assessments in December and May of Year 3, 7-years-old students, to assess their literacy competences in five different tasks – picture to word matching, sentence to picture, spelling, reading comprehension, and writing. Following a three-year cycle, it has been repeatedly noted that the students were performing poorly. In one state college, the results attained were discussed with the respective Heads of School and subsequently the class educators. The intent of these meetings was to further analyse the gaps and how the school and the college could best tackle them. What was certain was that an action needed to be taken by the primary schools forming part of the college in terms of training and most importantly the differentiation to be offered to these students. As a result, a training programme was designed and implemented along with two pilot projects – a school which tracked students according to their literacy levels for all the subjects like streaming on the basis of the literacy competences – and another school which used sets for the Maltese and English languages only. In monitoring the progress of both schools and based on the latest literacy assessment results, it is transpiring that the school working with sets is reaping results and hence, such findings will be presented by the school during the Council of Heads – a meeting with all the Heads of Schools forming part of the college – for possible adaptation at the other schools. With the college's mission to learn and grow together to be the best we can, we hope to reduce significantly the percentage of students leaving primary education with poor literacy skills.

# Symposium 2: Education Futures in Flux: Journeys into Learning & Assessment Transformation

16:00 - 16:20

## Assessment Futures Through a Looking Glass

B. Maddox[1]

[1]Digital Education Futures Initiative, Hughes Hall, University of Cambridge, United Kingdom

In the first presentation in this symposium I will present the results of a recent horizon scanning project [location/author anonymised] that explored digital assessment futures. Viewing assessment through the lens of futures thinking we identified current trends, signals, drivers of change, and wild cards. This presentation will describe some of the disruptions and changes that are likely to take place over the next decade. The surreal characters encountered on this journey include those supported by brain implants, augmented/immersive reality, stealth assessments, and large AI models. This paper considers their impact on, and relation to established assessment beliefs, educational values and standards. To do this, we will draw from the philosophy of technology, and in particular on the notion of the pharmakon (Stiegler, 2019, Simondon, 2023, Bateson, 1978), in which techniques create ambiguous ‘transitional objects’ that can be seen as a remedy, or a poison, where we need to distinguish between dreams and nightmares. In its conclusion, the paper proposes an evaluative framework, to think through those possible impacts and consequences, and to support value-based decision making about technology adoption.

# The Modern Socratic Assessor: The Promise of AI for the Future of Education

A. von Davier[1]

[1]ACT, USA

The recent launches of large multimodal computational models made us all dream of the modern Socratic tutor and assessor, who is accessible to all, of the immersive experience of learning and training, and of virtual multimodal assessments. In this presentation I'll focus on three areas of interest: 1. Construct definition: what and how should we teach in the times of AI? 2. Assessment design & development: what and how to assess the relevant skills in the New World (see von Davier, Mislevy, & Hao, 2021); 3. Social context: what type of guardrails do we need to protect and support students, teachers and integrity of the educational experience. I will present an example of the application of several AI tools (GPT3/4, BERT, voice generator) for generating a new item type, an integrated interactional competency task, the computational models for scoring this task, and I will discuss what types of AI Standards are necessary for assessment. I will conclude with my personal thoughts on the value of humanity in a fast world of technological prowess for education.

# Is this the future of essay writing? ChatGPT’s impact on the process and output in different languages

R. Hamer[1]

[1]International Baccalaureate, Netherlands

Being able to communicate clearly across cultural and language barriers is a skill that the International Baccalaureate (IB) values highly and features prominently in its summative assessments. The release of large language models (LLM) claiming to write texts indistinguishable from human writing means that current IB assessment tasks may require redesign to remain valid in the future, especially if there are unknown biases within the AI tools that result in unfairness across groups. This presentation focuses on the outcomes of using OpenAI and ChatGPT to generate over 150 samples of credible student work for existing high-stakes assessment tasks in three languages. Analysis of detailed process journals documenting the creation of each sample allows a comparison of the participants’ learning journeys and differences in AI performance across languages and tasks. While the task requirements are identical across languages, the ease of use of the LLM, its performance and the quality of the outcome in English seems to be superior to that in other languages. The existence of a significant language bias in current generative AI needs to be considered when adapting pedagogy and assessment design to the use of generative AI.

# Symposium 3: The Digital Transformational Journey in England: Lessons Learned from TIMSS 2019 and Implications for National Assessments

16:00 - 16:20

## A Comparison of the Delivery Considerations of Digital and Paper Assessments in TIMSS 2019 in England

M. Mohan[1], S. Turner[1], A. Hooper[1]

[1]Pearson, United Kingdom

Trends in International Mathematics and Science Study (TIMSS) 2019 marked the beginning of digital transformation by providing the option between paper and digital assessment modes to 64 participating countries. In England, TIMSS 2019 surveyed 9595 pupils across 368 schools, with 6761 pupils in 275 schools in digital mode, while 2834 pupils in 93 schools administered paper assessment. Digital assessments present operational delivery opportunities and challenges in schools and in post-assessment activities, such as scoring and data processing. Many schools face practical difficulties in assessing students digitally, from the availability of IT resources to the level of technological familiarity of students and teachers. Whilst there are advantages to scoring in digital assessments being taken onscreen, consideration is required in how to approach this important process and how it links to data processing.

Using the example of TIMSS 2019 in England, this paper provides details of the delivery of this assessment in schools, compares digital and paper assessments, and evaluates what can be done to overcome challenges across the dual modes before, during and after assessment delivery. Insights from this paper have potential to assist education systems in making evidence-based decisions on aspects of digital assessment delivery and post-assessment data activities.

# TIMSS 2019 Equivalence Study: A Quantitative Approach to Explore Assessment Mode Effects on Mathematics Performance in England

S. Nastuta[1], L. Liu[1]

[1]Pearson, United Kingdom

Using TIMSS England 2019 Grade 8 pupils' Mathematics achievement data, this paper compares 114 identical items delivered in both on-paper (N > 300) and on-screen (N > 400) format to identify any differences generated by delivery mode. This paper adopts a well-designed combination of quantitative analyses to explore the mode impact. The t-test of mean differences regarding pupils' final maths outcome shows that there is an overall assessment mode effect. England Grade 8 pupils found digital assessment slightly harder than paper assessment, but the difference is not statistically significant. Nevertheless, when looking at the item-level differences, the picture is more nuanced when checking the item discrimination and difficulty levels by IRT analysis for matched items across two modes. This paper contributes to the field by empirically showing that more granular research can provide insights into individual item-level gaps. Furthermore, the results of the correct response ratio for each question item in dual assessment modes found the questions requiring annotation, graphs, shapes, and visuals have the largest gaps, in line with the existing literature. The inherent difficulty of such items was also examined using process data (number of screens accessed and time spent on each item).

# The Impact of Assessment Mode on Item Performance: A Qualitative Study of TIMSS 2019 in England

K. Mason[1], L. Liu[1]

[1]Pearson, United Kingdom

This study investigates modal differences between digital and paper-based items in TIMSS 2019 in England following data analysis of correct response ratios of 114 identical items that appeared in both modes. Previous literature has explored modal effects mainly through experimental design or secondary data analyses. This paper therefore fills a research gap, using semi-structured interviews to gain insight from subject matter experts and TIMSS scorers to explain the impact of assessment mode on mathematics item performance.

The interviews investigate the advantages and disadvantages of each test mode, illustrated by seven items that showed substantial differences in the correct answer ratios across modes (>15%). Drilling down to the question item level, the qualitative analyses indicate that the nature of a question, including students' working load and exam-taking strategies, may vary between modes. For example, paper-based pupils benefit from the use of scratch paper to aid working memory. Furthermore, our qualitative results, in line with the literature, demonstrate questions requiring scrolling or opening a separate window are more subject to mode effects. Therefore, using an-inbuilt calculator and equation editor in digital TIMSS assessment may require practice, to improve pupils' familiarity and self-confidence during the assessment.

# Symposium 4: Supporting countries to set global standards on national learning assessments

16:00 - 16:20

## Learning Progression Scales

E. Stubbs[1], U. Schwantner[2]

[1]ACER UK, United Kingdom
[2]Australian Council for Educational Research, Australia

‘Learning Progression’ is a term used for a comprehensive description of what it typically looks like for learners to move from early through to advanced knowledge, skills and understandings within a learning area, such as reading or mathematics. We have developed empirical Learning Progression Scales (LPSs) in reading and mathematics, drawing on assessment items and data from a wide variety of programs used in international, regional and national assessment programs. The LPSs are an essential underpinning for our work supporting countries to set global benchmarks on their national learning assessments. The LPSs can also be used for curriculum review and assessment reform with the view of establishing where learners are at and what skills they need to progress.

This paper will explain the theoretical and contextual background to the LPSs and describe the steps in the development process.

# The International Standard Setting Exercise to locate global Minimum Proficiency Levels on the Learning Progression Scales

M. Walker[1]

[1]ACER, Australia

The global Minimum Proficiency Levels (MPLs) describe the minimum expectations for learners in reading and mathematics at important stages of schooling – the end of grades 2/3, at the end of primary and at the end of lower secondary. The MPLs were developed through an international consultation process and are used by UNESCO Institute for Statistics (UIS) to enable countries to report against Sustainable Development Goal 4.1.1 on the proportion of learners meeting these MPLs. The International Standard Setting Exercise was conducted to locate the MPLs on the Learning Progression Scales (LPSs) that we have developed to describe learning development in reading and mathematics.

This paper will explain the methods used to set the MPL benchmarks on the empirical LPSs, demonstrating the proof of concept that empirically derived 'long-span' LPSs can serve as the basis for locating MPLs on an empirical continuum. The paper will also explain the method used in this exercise, namely a process combining modified Angoff and bookmark methodology, and provide details on the range of international participants, procedures and instruments used.

# The Pairwise Comparison Method for linking national assessments to global standards

C. Watson[1]

[1]ACER UK, United Kingdom

Data on outcomes in the learning areas of reading and mathematics are central to monitoring and reporting countries' progress towards achieving UNESCO's Sustainable Development Goal indicator 4.1.1, by 2030. Such data is also essential to support countries on their assessment reform journey. Large-scale national learning assessments are widely recognised as a primary source for such data; however, they vary in method and scope, posing major challenges for global monitoring. Also, not all countries participate in international or regional assessment programs. Therefore, we have developed approaches to harmonise quantitative data across assessment programs, and to provide substantive information about children's learning levels and progress benchmarked against international standards.

This paper will provide details of the Pairwise Comparison Method (PCM), which is an approach making use of comparative judgement to set global benchmarks on national learning assessments. The PCM approach makes use of the empirical Learning Progression Scales that we have developed in reading and mathematics on which the Minimum Proficiency Levels have been set during an International Standard Setting Exercise.

# Symposium 5: Assessment reforms in Norway: tensions in quality assessment and quality development

16:00 - 16:20

## Tension points in assessment reform in Norway

T.S. Prøitz[1]

[1]University of South-Eastern Norway (USN), Norway

This paper presents and discuss the ongoing policy process of reconsideration and renewal of the Norwegian National Quality Assessment System. The paper aims to thematise how policy seem to underestimate the inbuilt tensions of quality assessment systems concerning the various and often conflicting needs of teachers, school leaders and educational leaders in municipals. In Norway, issues regarding the national quality assessment system including national tests, national examinations and student surveys have been raised several times over the years. However, the events of Covid-19 seem to have sparked the debate and the current government have taken several initiatives signalling new assessment directions for the future. Central policy ambitions are enhanced trust in professionals, fewer tests, less reporting and more focus on process and quality development in schools. Still, this is a contested terrain in policy as well as between practitioners at the different levels of the education system. Drawing on an analysis of assessment policy documents the inbuilt tension points of the old assessment system are identified and discussed in light of potential new policies.

# Quality assessment and -development: enactment through educational leadership autonomy

R.A. Sundberg[1]

[1]University of South-Eastern Norway (USN), Norway

The concept of quality in education is an important framing for assessment policy and practice. Recent educational reforms in Norway emphasize quality development processes, while quality assessment systems still to a large degree makes use of output results as quality markers. As global trends towards output-control of schools are often combined with extended local leadership autonomy, tensions between autonomy and control through quality assessment might arise, and school leaders' autonomy can be affected by several contradictory elements. Aiming to gain deeper understanding of which and how different elements inform the enactment of leaders' decision-making processes in quality assessment and -development, the study asks: Which elements inform or regulate decision-making processes related to quality assessment and -processes in local education leadership, and further what sources of tensions arise with recent assessment reforms in Norway? The study takes an ethnographic approach following leaders in two upper secondary schools in Norway during one school year. Preliminary findings indicate that leaders' decision-making in schools is particularly challenged by aspects of teachers' professional, social, and emotional needs as well as administrative structural elements. Conflicting elements rise to the surface when larger changes to practice are required, i.e. when implementing new curricula reforms or reorganising institutions.

# Validating oral examinations through a unitary view of validity

M.S. Syverud[1]

[1]University of South-Eastern Norway (USN), Norway

The implementation of the Knowledge Promotion Reform in Norway in 2006, led to increased focus on assessment in policy documents. The introduction of a new curriculum reform in 2020, however, has led to little change in the assessment system. Nonetheless there is ongoing debate between stakeholders about possible changes in the exam system; but this debate often lacks reference to empirical research, which is scarce within the Norwegian context. The aim of this paper is therefore to validate how oral exams are carried out in the subject Norwegian Language and Literature in four different secondary schools in Norway. The goal is gaining knowledge about validity issues tied to the practice of oral examinations by studying video-recordings of authentic oral exams. The research question guiding the study is: What sources of validity evidence can be found that either strengthen or weaken the argument for using oral exams for ranking students for further education? Applying Messick's unified view of validity this study considers different validity-related evidence and takes a qualitative approach to research on oral exams, including thematic analysis and content analysis. While expecting high validity, preliminary findings indicate that policy instructions encourage a practice affecting validity both negatively and positively.

# Symposium 6: Technicians, Curators or Guides on the Assessment Reform Journey? Preparing the Next Generation of Educational Measurement Professionals

16:00 - 16:20

## What are foundational competencies in educational measurement and why should we care about them?

D. Briggs[1]

[1]University of Colorado, USA

The first presentation will give an overview of the foundational competencies depicted in Figure 1 above, describing both what they entail, and how they interact. It will also make the case that this is a framework that people actively engaged in the educational assessment community should be prepared to discuss and debate. Some key principles behind the competencies will be elaborated. For example, that competencies are not intended to be exhaustive, but represent a subset of a fuller set that educational measurement professionals can possess and develop. They are also intended to be aspirational. This presentation will also provide some illustrative examples of how foundational competencies could be developed and nurtured though (a) curricular and extra-curricular activities that comprise masters and graduate programs, and (b) on the job training opportunities. It will conclude by raising the question of whether the NCME framework could (and/or should) become a basis for the certification of educational measurement professionals.

# Computational psychometrics skills in the age of artificial intelligence

A. von Davier[1]

[1]ACT, USA

The second presentation delves into the unprecedented impact of contemporary computational techniques and advancements in artificial intelligence on the field of educational assessment. It will begin with a description of computational psychometrics (CP) as applied to learning and assessment, consider the requisite skill set for professionals in this sphere, and then move to a discussion of the evolution of CP from a mere extension of the operationalization capabilities to a radical transformation of the underpinnings of educational measurement.
Furthermore, this presentation will explore the degree to which proficiencies traditionally deemed pertinent under the established framework may or may not persist in shaping the practices of the field. Illustrative case studies of the synergistic application of artificial intelligence and psychometrics in learning and testing programs will be provided.

# Too much or not enough: challenges of teaching foundational competencies in educational measurement

D.T. Irribarra[1]

[1]Pontificia Universidad Católica de Chile, Chile

The third presentation will tackle the practical challenges of considering a framework of minimal competencies as the basis for academics and professionals working in the field. This presentation will discuss the implementation challenges associated with developing these competencies in higher education institutions, and the demands that confidently mastering all perspectives in the framework can present to both new and experienced members of the educational assessment community.
It will begin by reviewing general lines of debate that have emerged from the community during the development of the framework, including calls for considering ethics and other areas in addition to the domains already proposed. The second part of the presentation will discuss the practical challenges that arise from expecting to address all the domains proposed in the framework in the context of higher education and, potentially, in the context of accreditation. Finally, the presentation will discuss recommendations on how the framework could be developed and applied at different levels, including undergraduate studies, graduate education, and professional development.

# Ignite Session

16:00 - 18:00

## The Key to Successful Assessment Reform: Authoring Reform

S. Crowley[1], A. Leigh[1]

[1]GradeMaker, United Kingdom

Assessment reform inevitably introduces new challenges for assessment authors and moderators.

Our instinct might be to avoid introducing new processes while new specifications are being phased in. But in this Ignite session, the presenter will argue that preparing for assessment reform is the best time to address authoring.

The presentation will focus on the benefits of introducing a specialist assessment authoring system to support reforms. It will explain how awarding organisations can achieve a smoother, more successful introduction of new specifications by:
- Improving the way assessments are mapped to subject taxonomies and performance criteria
- Improving the way authors are commissioned
- Providing specialist tools for development and quality assurance
- Streamlining the pre-testing process and making efficient use of performance data and other metadata
- Allowing for flexibility in delivery and design of assessments where reform is being phased over time or by subject or qualification.

# Development of a Game-Based Assessment of Divergent Thinking

L. Sun[1], Y. Yuan[2], F. Luo[2]

[1]University of Cambridge, United Kingdom
[2]Beijing Normal University, China

Divergent thinking tests measure the ability to generate multiple solutions to a given problem. They tend to be abstract, static, and unrealistic, resulting in low ecological validity. Human raters are usually required, rendering the scoring process subjective, labor-intensive, and time-consuming. To address these concerns, this study developed a video game consisting of three puzzle-like problems, where participants were allowed 45 minutes to identify possible solutions by selecting and ordering tools (i.e., potentially useful objects). The game was piloted by 515 undergraduate students. The results showed that the overall performance indicator, derived from a CFA model, effectively predicted the performance on the external measures of creativity. The three problems exhibited acceptable difficulty and internal consistency. 80.0% of the students found the game interesting. These findings provide initial evidence for the game to serve as a valid and reliable instrument to measure divergent thinking in a complex, dynamic, and interactive gaming environment.

# Partnership between schools and universities in developing assessment practices - the journey from intentions to enactment.

K. Haaland[1], H. Havn[2], E. Bræin[2], M. Rustad[2], C. Nygaard[2], L. Dahl[2]

[1]Innland University of Applied Sciences, Norway
[2]Inland University of Applied Sciences, Norway

In Norway, a new curriculum and assessment regulation has been introduced, where the focus on implementation by the Parliament has shifted from a centralized to a decentralized model. This involves a collaborative effort between counties and universities to jointly define, prioritize and develop teaching and assessment practices that support and promote student learning in schools.
Our presentation will highlight tensions and opportunities in a decentralized competence development model, in our experience as a partner representing the university.
We will focus on the tensions that arise when a new intention is introduced to professionals and interpreted individually and through prior knowledge.
We experience that the new concept of competence challenges the teaching and assessment practices of schools. We will discuss the tensions that arise in schools when professionals are asked to shift their focus from assessment of the student`s outcome, to the student`s understanding and reflection on the learning process.

# Creating the conditions for successful assessment reform through education system planning

B. Wyatt[1], M. Neesam[1]

[1]Cambridge University Press and Assessment, United Kingdom

An essential factor for successful assessment reform is continuing professional development for school leaders and teachers. There are numerous cases where assessment reform has caused disruption and unintended consequences because those implementing the reform at the school level of the system are not provided with the assessment literacy required to write effective assessment policies to align with the reform.

A major issue is the lack of focus on the key principles and purposes of assessment at the point of initial teacher education, leaving those entering teaching, and those within the system, feeling unprepared for large-scale change.

This presentation focuses on a strategy to support ministries with professional development to ensure there is alignment between the intended expectations of reform and the system operations, e.g. initial teacher education, needed to implement the change. This approach focuses on ministries creating the conditions for success and reducing the opportunity for unintended consequences.

# Why so many assessments? A holistic framework to help teachers to see the bigger picture - including the missing pieces

I. Suto[1], S. Crocker[1]

[1]Cambridge CEM, United Kingdom

Data can overwhelm teachers. There are numerous types of assessment, created by government bodies, assessment organisations and schools themselves. It can be challenging to understand the purposes of different tests and evaluations, the pedagogical questions they address, and how they can fit together coherently. Moreover, teachers may need to identify gaps and unwanted overlaps within their assessment approaches.

In this talk we present a teacher-friendly, evidence-based framework of five interacting areas of teacher insight into their learners' journeys. We argue almost all assessments, evaluations, and teaching resources can be understood in terms of these areas, and that our framework is a useful organising instrument for creating learner profiles. During face-to-face workshops in Malaysia, Indonesia, Thailand and Vietnam, we introduced the framework to 200 teachers in international schools. They applied it in group activities then gave feedback through questionnaires and plenary discussions. 86% reported the framework made a lot of sense.

# Views of Scottish disabled learners/learners with additional support needs (ASN) on National Qualifications assessment in 2022

S. Allan[1], M. Cuninghame[1]

[1]Scottish Qualifications Authority, United Kingdom

We researched the views of learners with experience of 2022 National Qualifications assessment. The research covered the reintroduction of external assessment in 2022 and those parts of the assessment process that had not returned to pre-pandemic approaches. We also asked learners to compare the 2021 and 2022 approaches.

We specifically analysed the views and experiences of disabled learners and/or learners with ASN. In some areas, the responses of disabled learners and/or learners with ASN were significantly different from the wider learner population. Understanding these results is essential to ensure fair and equitable reform of Scottish assessment and qualifications.

Disabled learners and/or learners with ASN were less likely to know details of the assessment process, such as around appeals, modifications to assessment and revision support. Additionally, disabled learners and/or learners with ASN found both the 2021 and 2022 assessment processes to be more stressful than learners as a whole.

## Perspectives of End-users and the General Public on Assessment II

9:00 - 9:30

### Setting Standards in the new Technical T Level qualifications in England: Prior attainment relationship to outcomes

J. Kaur[1], B. Ashworth[1]

[1]Pearson, United Kingdom

The Department for Education in England launched a new set of Technical Qualifications called T Levels in 2020 to meet the needs of industry and employers in England and following a Post 16 Skills report identifying a potential need to improve the technical/vocational path to adequately equip students with the skills needed for the workplace.
These are post 16 qualifications primarily designed to enable students to progress into employment, however a proportion of students completing these qualifications use them as a progression route into Higher Education. Setting and maintaining standards for reformed qualifications is challenging and setting standards for completely brand-new qualifications such as T Levels bring a whole array of new questions and challenges, especially with the blend of both the academic and vocational content.
As T Level qualifications have some assessments focused on assessing knowledge, exploration into whether methodologies for maintenance of standards as used for academic A-levels is considered alongside exploring the potential impact of adopting academic methodologies could have on the trust and public confidence in these technical qualifications given their purpose to demonstrate a level of competence.

# Spoken practice: evaluating an external ESOL language assessment for young learners in France

J. Champaud[1], B. Seguis[2]

[1]Cambridge University Press & Assessment, United Kingdom
[2]Cambridge University Press and Assessment, United Kingdom

This paper will present key findings from an impact evaluation of a school-based English learning and assessment programme in three state sector primary schools of a town north of Paris.

Evaluating the impact of these external inputs was of added salience due to its reliance on specific local budgetary priorities, and a backdrop of learning and assessment reform in France for English as a foreign language in the 1st and 2nd cycles.

The study employs a mixed-method methodology, triangulating assessment scores, with survey data. To capture a more qualitative and nuanced picture of the impact on classroom practices, school policies, and community engagement, we conducted interviews with key informants and educators.

The findings showed positive outcomes on learners language proficiency and motivation, as well as parents and teachers' confidence and attitude towards English language assessment. More importantly, the study demonstrated the positive impacts of the project on stakeholders' attitudes towards external assessment and schools ability to foster - through the project - an immersive and enabling culture of English learning and assessment. The study discusses the implications of this case study on future policy and assessment reform for foreign language acquisition in France.

# The Adaptive Models of the New National Screening Tests in Reading and Numeracy for Grade 1 and 3 students in Norway - possibilities and limitations.

B. Walgermo[1], G.A. Nortvedt[2], P.H. Uppstad[1], K.B. Bratting[3], H.H. Haram[2], N. Foldnes[1]

[1]University of Stavanger, Norway
[2]University of Oslo, Norway
[3]Universitetet i Oslo, Norway

The assessment field has witnessed an increased interest for more personalized adaptive tests. This quest is of the highest relevance in the field of special and inclusive education when developing tools to identify students in need of extra follow up, so-called screening tests. Yet, information on the development of such adaptive screening tests is scarce. The present study describes and contrasts the development of two new screening tests in numeracy and reading, for 1st and 3rd grade, in Norway. These tests are built on two different adaptive routing models. Through student observation, teacher surveys and interviews we have also investigated the interpretation and use of the new screening tools in numeracy and reading. Put shortly, teachers report the tasks to be meaningful and engaging for the students. However, teachers describe test length to be the most important challenge to overcome in order to optimize test experience for the young struggling students. Strengths as well as suggestions for improvements in the two adaptive models for better user experiences, including possible approaches for reduced test length, will be discussed.

## Summative Assessment II

9:00 - 9:30

### Assessing Knowledge Acquisition through Brain Activity: Shared Processes of Active Learning Strategies in Vocabulary and Mathematics?

B. Jonsson[1]

[1]Umeå University, Sweden

Research has compared active and passive learning strategies in different subjects such as mathematics and vocabulary learning, across a wide range of subjects, cognitive abilities and age groups. Creative Mathematical Reasoning (CMR) has been developed for mathematics to emphasize problem-solving. While retrieval practice (RP) has been used for actively recalling information from memory during vocabulary learning. A recent within-subject study using functional Magnetic Resonance Imaging investigated whether active learning using CMR and RP engages similar learning processes across the different subjects, mathematics and vocabulary, respectively. The study revealed a shared brain network when students were tested one week after the learning session in regions such as the precuneus, inferior parietal cortex/angular gyrus, and left lateral and medial prefrontal cortex. This suggests that the benefits of active learning may result from the recruitment of overlapping learning processes and networks in the brain, independent of learning strategy. The study provides support for the hypothesis of engagement of a shared brain network at retrieval after active learning of mathematics and vocabulary. It also raises the question, what do we assess when we assess learning methods?

# Accessibility in high stakes testing and validation of test accommodations: Empowering visually impaired students.

M. Strömbäck Hjärne[1, 2], C. Wikstrom[3]

[1]Luleå Technical University, Sweden
[2]Umeå University, Sweden
[3]Umea University, Sweden

In fair testing, all test takers should be given the same possibility to perform at their best. However, although test accommodations are common, perhaps especially in large scale testing, it is an under-researched area with very little practical guidance. This research project aims to increase accessibility to the Swedish admissions test to higher education, the SweSAT, for visually impaired students. The intention is also to propose a theoretical and practical framework for validating accommodations in a broader perspective. The project implements well-recognized validity theory, utilizing a central validity argument approach to assess adaptations and accommodations made for visually impaired students. The project also seeks to address potential tensions in the process of implementing accessibility measures, and the impact on various stakeholders. Our findings will contribute with valuable insights into the challenges and opportunities of creating accessible and equitable assessment environments with potential to inform future assessment reform initiatives.

# Defining standards in a reformed national qualification system: lessons in coherence from Covid to recovery

R. Harry[1]

[1]WJEC, United Kingdom

Assessment systems have many features, including what performance evidence is produced by students, the conditions in which it is produced, who determines a student's overall outcome, how different types of evidence are combined to determine that outcome, and how standards are defined. Assessment reform can target some or all of these features to achieve its objectives, and changes are usually made carefully.

Between 2017 and 2019, reformed versions of the qualifications used by students in England and Wales - GCSEs and A levels - were completed for the first time. That period, and the period since, has seen unprecedented turbulence in these features, which were adjusted and reframed several times in the return to 'normality' this summer. This provides an unexpected but invaluable opportunity to learn how different features of an assessment system can be combined together to create a coherent assessment system.

Using the impending reform of GCSEs in Wales as a frame, this paper attempts to make sense of this turbulence, focusing in particular on the inferences made between performance and attainment, as a means to inform future assessment reform. The aim is to develop foresight, pre-empt issues and avoid the pitfalls that bedevil many reform programmes.

# National Tests & Examinations lll

9:00 - 9:30

## Performance in secondary mathematics topics pre- and post-reform

J. Williamson[1], C. Vidal Rodeiro[1]

[1]Cambridge University Press & Assessment, United Kingdom

In England, a recent major programme of assessment reform replaced the General Certificate of Secondary Education (GCSE) qualifications that young people take aged 16. In mathematics, the stated aims of reform were ambitious: to ensure mastery of fundamental mathematics by all students, while also creating a more challenging qualification to improve preparation for further mathematical study and careers.

Early qualitative research with teachers identified perceptions of how GCSE mathematics reform had affected teaching, learning and attainment. To better understand the reform's impact, we decided to complement these studies by offering a quantitative analysis of student performance in GCSE mathematics examinations. Our research analysed the performance in maths items of approximately 250,000 candidates from the final three years of pre-reform GCSE mathematics (2014-2016) and the first three years of the post-reform GCSE (2017-2019), focusing on performance in different mathematics topics and how this changed (or did not change). The findings confirmed that candidates found post-reform GCSE assessments substantially more challenging. The proportion of marks achieved decreased more in some topics than others, but the variation across topics was not statistically significant. It may nevertheless have affected teacher and candidate experiences of different mathematics topics, and perceptions of the reformed qualification.

# How to go about “eating an elephant”: A critical analysis of validity frameworks in application

P.J.(. Ho[1]

[1]University of Oxford, United Kingdom

Validation is known to be a strenuous endeavour in that it is resource-intensive and perpetually ongoing. The recency of the development of practical validation approaches adds to the challenge, as validators look for guidance on envisaging and conducting validation research.

This paper examines the strengths and limitations of three commonly referenced validation frameworks in evaluating the Hong Kong Diploma of Secondary Education Examination (HKDSE). The frameworks are the ‘five sources’ framework from the Standards (AERA et al., 2014), Newton’s (2016) macro-/micro-validation framework, and the chain model by Crooks et al. (1996). An analysis of the strengths and limitations of the frameworks in the context of the HKDSE is presented, followed by personal reflection on navigating the corpus of evidence and analysis as an independent researcher.

This paper primarily presents two arguments. Firstly, despite the similarities across the frameworks which share a common goal of increasing accessibility for validators, the choice of framework has a decisive impact on what may or may not be included as evidence and subsequently the strength of the validation argument. Secondly, the choice of framework should be determined within the context of the assessment procedure and purpose, specifically the paradigm of assessment it falls under.

# Towards a fairer and more equitable national test system - focusing standard setting and equating

A. Lind Pantzare[1]

[1]Umea university, Sweden

The Swedish national testing system is currently undergoing a major reform. It is a reform where validity is a guiding star with a goal to, in a couple of years, have tests that are digital, more reliable, effective and, fairer to the test takers. One issue related to fairness and equity concerns the process of setting cut scores for different test grades and whether that process provides equivalent requirements over time. Which test form a test taker receives should not affect the result.

In this study the cut scores from the regular standard-setting before test administration is validated through a comparison with equated cut scores after the test administration to answer the question if the proposed cut scores are fair and equal. The data consists of a random sample of student results from the regular administration of three consecutive administrations of the national tests in one of the higher courses in mathematics for upper secondary school. The results from the comparison of the cut scores from the standard-setting in relation to the equated levels will be presented and discussed.

# Formative Assessment III

9:00 - 9:30

## Exploring the Potential of Using Lesson Study to Enhance the Formative Dimension of Classroom Assessment

M.A. Buhagiar[1]

[1]University of Malta, Malta

Assessment reforms do not necessarily translate into tangible results. The evidence suggests, for instance, that classroom assessment reform policies are likely to fail unless they are planned and implemented with sensitivity to the embedding contexts. It thus appears that the way forward calls for initiatives that are closely attached to the space within which classroom assessment occurs, basically the classroom itself. On this basis, the presentation, which is conceptual in nature, seeks to explore how lesson study, which is intrinsically linked to what happens inside classrooms, could be used by teachers and their students as a tool to enhance the formative dimension of classroom assessment. This presentation draws on current understandings of formative classroom assessment and transposes the application of lesson study to classroom scenarios in which teachers and students join forces to plan, implement and evaluate ‘research lessons’. The aim of such lessons is to facilitate the ways of how teachers and students learn about student learning, and how they could then act on this learning in a way that leads to improved future learning. Ultimately, the presentation proposes a theory-driven model of how lesson study could be adopted to improve the formative dimension of classroom assessment.

# A teacher’s attempt to enact the vision of implementing formative assessment: A case study

T. Palm[1]

[1]Umeå University, Sweden

A teacher’s attempt to enact the vision of implementing formative assessment: A case study

Research has shown that classroom practices that adhere to the principles of formative assessment (FA) can accomplish large gains in student achievement, and FA has been promoted in many countries.

However, several issues make it difficult for teachers to develop FA. This study focuses on the obstacle that accomplishing high-quality FA is complex and difficult for teachers. We have studied one mathematics teacher’s FA when responding to students’ requests for help when they are working individually with mathematics tasks. The study is delimited to the teacher’s role as proactive agent in the FA processes. In this context, these FA processes are the teacher’s (1) elicitation of evidence of student knowledge and skills, (2) interpretations of the elicited evidence and the inferences made about student learning needs, and (3) feedback adapted to meet these learning needs.

The teacher audio-recorded her help to the students and sent the recordings to the researchers who analysed the recordings and provided feedback. Each such cycle lasted one month, and the procedure continued for one year. The difficulties the teacher encountered and how some of the difficulties were overcome will be presented.

# Assessment of Practical Skills II

9:00 - 9:30

## Assessing Practical Skills in High Stakes Qualifications: A qualitative study of GCSE Science qualifications

A. Hooper[1], C. Harrison[1], G. Grima[2]

[1]Pearson, United Kingdom
[2]Pearson UK, United Kingdom

In 2016, the reformed GCSE Science qualifications (high-stakes examinations usually taken aged 16), were launched in England, with first summative assessment in 2018. They moved to a linear assessment model, with fixed rules in place for practical assessments; 15% of exam marks related to practical work and schools to confirm students had opportunities to complete at least 8 practical activities during their course.
The literature suggests that the benefits of scientific practical work are wide-ranging, including increase in motivation and engagement with science. The aim of this study, completed in October 2022, was to evaluate the impact of resources on teacher and students' assessment experiences in the reformed GCSE Science qualifications, with a particular focus on the assessment of practical skills. Using semi-structured interviews with Heads of Science and Science Teachers, lesson observations, student focus group, and student survey, from seven secondary schools, the study looked to understand the experience of those delivering, and studying, a reformed science qualification in a post-pandemic era.
The key findings from the research were the impact of 'teaching to the exam', plus catching up on 'lost-learning' from the pandemic, on the delivery of practical science lessons across Key Stage 4 (14-16 years old).

# The impact of using peer assessment in writing essays

N. Orazbayeva[1], A. Kelimberdyieva[1]

[1]Nazarbayev Intellectual school, Kazakhstan

The main aim of this study is to investigate the impact of peer-feedback in a face-to-face classroom according to EFL learners' writing skill. This experimental study was conducted using through questionnaire, interview, pretest and posttest scores, experimental and control groups, each including twelve EFL learners in each group, were selected as a participants of the research. To assess the learners writing skills, researchers utilized writing IELTS task 2. Identifying problems, convincing arguments, clarity of statement, coherence and lexical resources, grammatical accuracy are assessed throughout the peer assessment process. They were taught the structure of problem solution essay and cause and effect essay, and were provided with the rubric, providing written comments in order to evaluate fellow students' writing performance and develop their writing skills as an assessor. The quantitative features taking from pretest and posttest were analyzed thoroughly with pre-questionnaire and post-interviewing. The results demonstrated that peer feedback accompanied the students (72%) to boost their writing skill and to develop their self-sufficiency. These results lead to the conclusion that peer feedback can be a useful technique for enhancing students' peer modifications in second draft, and may also improve EFL learner's performance on a written post-test.

# A multi-country comparison of lower secondary students' critical thinking under different curricula

J. McGrane[1,2], S. Johnston[3], M. Vendrell i Morancho[4], A.T.N. Hopfenbeck[1,2]

[1]University of Melbourne, Australia
[2]Kellogg College, University of Oxford, United Kingdom
[3]The University of Oxford, United Kingdom
[4]Complutense University of Madrid, Spain

Critical thinking is an essential skill for life-long learning, and given its increasing importance as a graduate attribute, it is vital to robustly evaluate how systems can best improve students' critical thinking. This study evaluates the differences in the critical thinking skills of students in the International Baccalaureate (IB) Middle Years Programme (MYP) versus students enrolled in the national curricula in Australia, England and Norway. The study comprised of 870 MYP (n = 386) and non-MYP (n = 484) students in Grades 9 and 10 across 21 schools. Data were remotely collected on their critical thinking skills and several other relevant cognitive, non-cognitive and background characteristics. A propensity score matching approach was used to match the MYP and non-MYP groups on these characteristics and they were compared using weighted regression. Findings showed that, overall, IB MYP students possessed significantly higher levels of critical thinking skills than their non-IB MYP peers with a moderate effect size. This advantage also held at both grade levels and across Australian and English students, with no significant difference for Norwegian students. Thus, the MYP appears to be a promising exemplar for enhancing critical thinking. Based on these findings, recommendations for critical-thinking pedagogy are offered.

## Fairness & Social Justice III

9:00 - 9:30

# How should we design central tests to ensure they are universal? Different perspectives on the use of accommodations in centralized testing in Flanders

S. Dierick[1, 2], P.d.K. Struyven[2]

[1]kuleuven, Belgium
[2]UHasselt, Belgium

The best way to achieve inclusive assessment is to start from the principle of universal design for assessment. This study aims to investigate different perspectives on the use of accommodations in centralized testing with the purpose of discovering (1) which accommodations are desirable for students with special educational needs (SEN) and what conditions are attached when granting accommodations? (2) How universal should we design central tests? The data were collected via focus group discussions (n=20), organized via educational partners (n=182) with different perspectives. This was followed by a thematic analysis and a data-driven approach for the coding methodology. The results show that most of the participants agree that students with SEN should be allowed to use the accommodations they are familiar with in class. Regarding universal design, the majority indicate that as many tools as possible should be made available to all students. Nevertheless, there are important tensions and concerns formulated from the different perspectives. The results of the study contribute to the debate of how best to create an inclusive testing environment in central tests that does justice to all students, tests the construct with as few barriers as possible, and maintains the validity, reliability of large scale testing.

# Examining the domain relevance of a test that has differing designated purposes- Is it valid?

L. Miller[1], R. Clesham[2]

[1]Pearson, United Kingdom
[2]Pearson UK (corporate membership), United Kingdom

Global high-stakes English language proficiency tests were initially purposed to assess the language proficiency of test-takers entering higher education in English speaking countries. Increasingly these tests are also accepted as part of the entry requirements for skilled economic migration into the same English-speaking countries. Visa entry requirements are equally high stakes and can change the course of the lives of individuals and their families.
Language proficiency tests do fulfil the main purposes of an assessment system, however, in terms of assessment validity, it is important to conduct validation studies to support each new test use context.
This session describes such a validation study in terms of the relevance of one English language proficiency test to skilled migration entry to Australia. The study engaged with stakeholders responsible for the standards of proficiency required by their professional field, well acquainted with the diversity of these roles and skills in Australia. Unusually, performance descriptors of these visa categories do not exist, only test score requirements. Therefore, as part of this study, performance descriptors were also developed for the visa categories, rooted in the requirements of professional bodies and based on test taker and test item exemplification across the four skills of language proficiency.

# We need to talk about SEND: How can a needs-based approach to assessment design result in fairer assessments for learners with Special Educational Needs and Disabilities?

I. Custodio[1], E. Barrow[1], D. McVeigh[1]

[1]Pearson, United Kingdom

More inclusive digital assessments provide opportunities to accommodate diverse student needs. Understanding these diverse needs is critical for the development of accessible and valid digital assessments. Much of the existing assessment research literature has treated Special Educational Needs and Disabilities (SEND) as a homogeneous category. In this presentation, we report on findings from a literature review that explores moderate to severe SEND conditions and implications for digital assessment design.

This literature review provides an important insight into the assessment requirements for specific SEND groups. Focusing on five conditions, we consider the characteristics of each SEND condition, and the likely implications for different modes of assessment. Recommendations explore not only how students 'access' assessment, but how the overall design enables them to demonstrate their full abilities.

The findings and recommendations from the literature review suggest that to address the complexities of needs across different SEND groups, benefits may be better realised by a needs-based approach, rather than through a solely condition-focused lens. We argue that a homogenous definition of SEND is limiting within the current educational landscape, and a better recognition of the diversity of and within SEND conditions is paramount, particularly as we move towards an increasingly digital education system.

# Psychometrics & Test Develpoment IV

9:00 - 9:30

## Using selected response items to assess higher order thinking skills

E. Sweiry[1], M. Hodgkin[1], L. Kennedy[1], A. Loomes[1]

[1]Ofqual, United Kingdom

Compared with constructed response items, selected response (SR) items assess a broader portion of the domain in a given period, are scored quickly and cheaply, and elicit very high scoring reliability. Nevertheless, SR items are viewed negatively in many contexts. One reason for this is a perception that they are less effective at assessing higher-order thinking skills (HOTS), which are typically defined as skills requiring levels of cognitive processing beyond knowledge and understanding. In this study, six item writers in three subjects (biology, psychology and business) wrote a total of 300 SR questions, targeted at 16 and 18 year-olds, designed to assess HOTS. They used seven different SR formats including multiple choice, cloze and matching. The items were reviewed by subject experts using a coding frame, and interviews were carried out with the item writers. A number of aspects were explored, including the effectiveness of the items at assessing HOTS (and particularly analysis and evaluation), the SR formats most conducive to assessing HOTS, optimal approaches to writing the items and the challenges faced in writing them. Assessment design decisions are the subject of complex balances and trade-offs, and the choice of item format is a significant factor in these decisions.

# The Application of Generalizability Theory to the Scenario-Based Performance Assessment of 21st Century Skills: Analysis of Task Context Effect

D. Gracheva[1]

[1]National Research University Higher School of Economics, Russia

The assessment of 21st century skills requires new test formats based on observed student behaviour in an interactive environment, such as Scenario-Based Performance Tasks (SBPTs). This paper aims to analyse the effect of the context of scenario tasks in measuring critical thinking and communication. The paper uses the methods of Generalizability Theory, which makes it possible to analyse to what extent the results can be generalised to other scenario task contexts and how reliable the assessment would be with different numbers of task contexts. The study is based on data from more than 2000 primary school students who were tested with different SBPTs developed within the Evidence-Centered Design framework. The results of the analysis showed that test takers' behaviour differed in scenarios with different contexts, while the main effects of the contexts were almost the same. To achieve satisfactory reliability, it is recommended to use at least two scenarios with different contexts. The study also assessed the effect of context when alternative forms of SBPTs were used. Overall, the effect of context should be taken into account when designing scenario tasks in order to obtain reliable results.

# Experienced but detached from reality: Theorizing and operationalizing the relationship between experience and rater effects

D. Tsagari[1], I. Lamprianou[2], N. Kyriakou[3]

[1]Oslo Metropolitan University, Norway
[2]University of Cyprus, Cyprus
[3]Frederick University, Cyprus

It is often argued that, to achieve a high quality of rating in writing assessment, it is important to nurture a comprehensive measurement ecosystem involving experienced raters and appropriate rating scales. But what is 'experience', and how do different kinds of experience affect the severity and the reliability of rating? We collected data from a high-stakes English writing examination to investigate how 18 raters with temporally and qualitatively different experiences interpret and use a rating scale. We analyzed the data using mixed methods, including summative content analysis, Rasch models and Graph theory showing that qualitatively different experiences affect severity in different ways. Also, irrespective of past experience, raters who disengaged from their Community of Practice (CoP), were more likely to yield unreliable ratings. These results have important methodological implications in how researchers theorize and operationalize 'rating experience'. Our findings highlight the importance of active and uninterrupted engagement in raters' CoP with implications for both researchers and policy makers.

## Other II

9:00 - 9:30

# Evidence for assessment reform journeys: The Analysis of National Learning Assessment Systems

U. Schwantner[1]

[1]Australian Council for Educational Research, Australia

Initiated by the Global Partnership for Education (GPE) in 2018 and developed by the Australian Council for Educational Research (ACER), the Analysis of National Learning Assessment Systems (ANLAS) provides a resource for education systems to systematically gather and analyse information about their national learning assessment systems. Such analysis provides essential evidence for a holistic assessment reform that is highly contextualised and locally relevant.
The ANLAS model focuses on interlinkages between the assessment and education system across three dimensions: 1) Contexts (legislation and policy, institutional arrangements, governance structures, funding, and leadership); 2) Quality of assessment programs (large-scale assessment, examinations, and classroom assessment); 3) Coherence (with learning standards and curriculum, education system structure, or national education priorities).
A detailed manual and toolkit were developed to support implementation and guide the analysis. These tools are adapted to fit national contexts – ensuring the identified improvements are relevant and appropriate. The country-led, participatory process of ANLAS creates broad buy-in from stakeholders for assessment reform.
This paper describes the ANLAS model and process demonstrating how findings from three pilots in Ethiopia, Mauritania, and Vietnam in 2019 fed into education planning. Using retrospective methods, the sustainability of the initiated assessment reform processes is investigated.

# Assessment practice: pivotal in understanding and development of good learning cultures.

I. Jacobsen[1], K. Blichfeldt[1]

[1]Centre for Lifelong Learning, Faculty of Education, Inland Norway University of Applied Science, Norway

In 2020 Norway's assessment regulation was reformed, followed by a new curriculum. The new assessment regulations specifically emphasise that assessment should be integrated in the learning process and contribute to the "desire to learn." This paper describes ongoing partnerships with two larger municipalities in Norway. Beyond the practice of assessment, we illuminate the key role assessment plays in developing schools on several levels.

Research topic: How assessment plays a key role in understanding, exploring, challenging and developing a school's culture. We propose that examining and challenging the assessment culture in schools reveals the values underpinning teacher practice and the didactical choices teachers make.

Methodology design: The paper is a result of an explorative study, with formative interventions and action-based learning. The material is based on our experiences as partners in school-based development.

Theoretical framework: We use activity theory as theoretical optics. The activity system, a visualization of the theory and contradiction matrix, is used as a systematic approach to analyze the data and discover tensions.

Preliminary findings: One of the preliminary findings show that teachers are uncertain about how to interpret the regulations. We hope that our paper can inspire and provide different perspectives in how assessment reforms impact schools.

# An Assessment Framework for Education Reform Projects

M. Dean[1], D. Bray[1]

[1]Cambridge Assessment International Education, United Kingdom

For the last 15 years, our professional education organisation has collaborated on projects with various ministries, school groups and donor organisations on a wide range of assessment-related reforms.

In this presentation we will present an assessment framework that we have developed for such reform work, outlining and giving the rationale for the levels and sub-levels, and we will detail which specific activities we would expect to take place for each. We will show how the framework can be operationalised and adapted to a specific context and the requirements of the partner organisation. We will give examples of specific processes, documents and deliverables that have been used or produced when working with partner organisations.

We will also discuss the key features in terms of collaboration that project evaluation has shown is essential to a successful intervention, as well as some of the challenges that we have encountered when working on different projects globally. We will outline how the framework can be presented and socialised and consider the factors that may ultimately dictate whether the reform in question is successful.

# Keynote Speech

11:00 - 11:45

## Lost and Found: Navigating the Maze of Instructional Feedback

A. Lipnevich[1]

[1]Queens College & the Graduate Center, City University of New York, USA

In this presentation, I will describe a series of studies that have investigated instructional feedback, exploring its mechanisms and the diverse (and often paradoxical) effects it has on various educational outcomes. I will delve into the underlying processes that contribute to the effectiveness of instructional feedback and discuss conditions that optimize its potential for enhancing student performance, learning, and individual characteristics. I will share research on assessment approaches employed by instructors and highlight cognitive biases that may influence assessment-related decisions made by both teachers and students.

Further, I will propose strategies for equipping students with the necessary tools to generate self-feedback effectively, promoting autonomous learning.

The studies discussed in this talk will encompass a wide range of contexts, cultures, and academic disciplines, emphasizing potential pitfalls with generalizations of findings. Additionally, I will outline potential avenues for future research and highlight some of the current challenges faced by the field. By addressing these challenges, we can advance our understanding of instructional feedback and its implications for educational settings.

## Keynote Speech

11:45 - 12:30

# Many rivers to cross? Navigating the challenging terrains of assessment in education

M. Richardson[1]

[1]UCL Institute of Education, United Kingdom

The one continuous factor in most education systems around the world seems to be change of one kind or another. Such changes might be invoked by decisions of a political, social, national, local, or even personal nature, but they share a common feature – an impact, somewhere, on the lived experience of students and their teachers. Change is not something that humans seem to like and when it affects a critical aspect of our lives, as educational assessment is, the anxiety and concerns about its value are brought into sharp focus. However, it doesn't need to be this stressful and in this talk I will consider how we, as an assessment community:

should plan effectively for change in our practice
should create networks to share practice and improve understanding of assessment
should build bridges that invoke trust within and between those networks

As the philosopher Thomas Paine said "We have it in our power to begin the world over again" and we have the expertise and knowledge as a community to enact substantive and valuable change – not necessarily to begin again, but to build alternatives to systems that don't benefit individuals, societies and nations. We can take this challenge and use it to continue to improve the role of assessment in education.