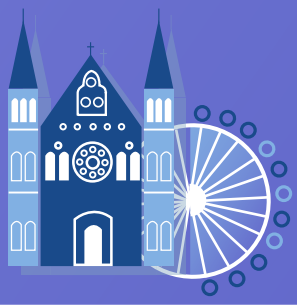
aea '25 EUROPE



05-08 November
The Hague, Netherlands

BOOK OF ABSTRACTS

Workshop 1

9:00 - 16:15

Identifying and analysing evidence to determine whether tasks elicit the intended constructs: bridging the gap between modern validity theory and innovative validation practice.

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

[1]Institute of Education, University College London, United Kingdom
[2]CIEA, United Kingdom
[3]Ofqual, United Kingdom

Establishing that assessment tasks elicit performances that reflect the intended constructs is a fundamental component of assessment validation. However, literature on this question is dominated by theoretical perspectives, with little practical guidance for practitioners. In addition, recent developments in the assessment landscape, including the emergence of technological advances (e.g. process data) and novel forms of assessment (including 21st-century skills such as collaboration and self-reflection), mean that established guidance may not adequately address contemporary requirements.

Through a combination of presentations, discussion and group activities, this workshop explores the challenges in identifying and analysing validation evidence to determine whether assessment tasks elicit performances that reflect the intended constructs. Participants will explore a range of validation evidence sources, from traditional statistical and qualitative approaches to innovative methods like process data and eye-tracking. Attendees will work through real-world scenarios, critically examining the alignment between task design and cognitive processes, and considering the implications of emerging technologies on validation practices.

Participants will leave with a deeper understanding of how to strengthen their validation arguments, navigate trade-offs in assessment design, and leverage both classic and emerging techniques to enhance validity in diverse educational contexts.

Workshop 2

9:00 - 16:15

Exploring feedback dialogues for a transformative feedback culture

S. Passeport[1, 2], A. Watts[3], M. Talbot[4], N. Younès[5], C. Höpfner[2]

[1]No Borders Learning, Netherlands
[2]Tilburg University, Netherlands
[3]University of Cambridge, UK, United Kingdom
[4]University of Leeds, United Kingdom
[5]Université Clermont-Auvergne, France

This interactive pre-conference workshop invites participants to critically engage with formative assessment or Assessment for Learning (AfL) by focusing on the articulation and shared understanding of feedback. Research highlights feedback as a powerful intervention (Hattie, 2009; Hattie & Timperley, 2007), yet its impact varies. Some approaches view feedback as information transfer, while others see it as a process. Building on the latter, this session examines feedback dialogues through a socio-material lens, exploring how relationships, power dynamics, tools, technologies, and institutional structures shape feedback encounters. Participants will engage with key research on feedback literacies, gaining insights into how feedback cultures are enacted and experienced by students and assessors. Through discussions and activities, they will reflect on their feedback experiences and explore strategies for fostering meaningful, dialogue-based interactions. Designed for educational researchers and practitioners—including teachers, lecturers, and curriculum coordinators in secondary and higher education—this session requires no prior expertise, just curiosity and an interest in student agency and dialogic feedback.
By the end, participants will have a deeper understanding of formative feedback as a relational, situated practice and leave with practical strategies to cultivate sustainable, dialogue-rich assessment cultures.

Workshop 3

9:00 - 16:15

Your best friend the psychometrician: The preventive role of psychometrics in test development

M. Van Onna[1], B. Hemker[1], C. Sluijter[2]

[1]Cito, Netherlands
[2]SCIEM, Netherlands

This workshop will help you to get a further insight in the advantages of timely involvement of a psychometrician when setting up a new testing program. It is useful for non-psychometricians to find out on how many more issues they can call on their friendly neighbourhood psychometrician. For psychometricians, the workshop may help to increase their added value.
During the workshop, we will use a scheme of all activities involved in test development. We'll discuss several general psychometric topics, and relate these to the decisions you will have to make for these activities. We'll show in which way a psychometrician might contribute to each activity. In each block, we'll give guidelines and illustrate best practices. We'll invite you to share your experiences with the topics and ask us for advice.
No R, no formulas, still all psychometrics.

Workshop 4

9:00 - 16:15

From awareness to action: embedding inclusive assessment in teacher development programs in higher education

C. van der Lienden[1], L. Koster[2]

[1]Risbo. Erasmus University Rotterdam, Netherlands
[2]Erasmus University Rotterdam, Netherlands

In today's diverse learning environments, inclusive assessment is essential to ensure fair and valid learning outcomes. Moreover, assessment should support higher education students in their learning process, enabling them to demonstrate their knowledge and skills. Inclusive assessment practices align with broader educational values such as contributing to a more accessible and equitable society.

To achieve such inclusive assessment practices, it is crucial that educators are supported in developing the necessary awareness, knowledge, and skills. Teacher development programs play a key role in this process. In the Netherlands, structured development trajectories are used to build assessment literacy in among university teaching staff, with inclusive assessment increasingly integrated as a central theme.

This workshop focuses on how inclusive assessment can be meaningfully integrated into teacher development programs in higher education. Led by an educational measurement expert and an inclusive education specialist from Risbo (Erasmus University Rotterdam), we take an evidence-based approach by translating research into practical strategies for program and course design, assessment construction, and grading. Through interactive activities and peer dialogue, participants will explore how inclusive assessment can be implemented in their own context, with concrete tools that contribute to fairer, more valid assessment practices and improved student learning outcomes.

Workshop 5

9:00 - 16:15

Establishing valid qualification equivalency with qualitative judgement

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

[1]Cambridge University Press and Assessment, United Kingdom

Where statistical equating methods are not available, through lack of common items or common candidates, but equivalency between two qualifications is required, it can be difficult to provide robust evidence.
Session 1 of the workshop focuses on a novel standard setting methodology that allowed for IGCSE scores to be translated into Mississippi end-of-course performance levels and integrated into the state accountability system. The method draws on aspects of both Body of Work (BoW) and Bookmarking methods to create an operationally feasible process.
Section 2 explores why there is a need for qualification equivalency, how a qualification can be broken down into content, demand and awarding standards, and some possible methodologies for establishing standards equivalency when the data for psychometric equating is not available.
The practical activity will involve delegates becoming comfortable with using the CRAS framework to evaluate the demand of questions, and how to set up, run and evaluate a comparative study using the No More Marking platform. We will discuss the benefits and limitations of this methodology for establishing demand equivalency, and what follow up work can usefully done with the results.

Workshop 6

9:00 - 16:15

Network Analysis for the investigation of Rater Effects (using R)

I. Lamprianou[1]

[1]University of Cyprus, Cyprus

This workshop introduces the application of Network Analysis (NA) to rater-mediated assessments. NA analyzes rating datasets by considering pairwise comparisons between raters.
Participants will learn how to detect and interpret key rater behaviors, including severity/leniency, inconsistency (misfit), halo effects, bias, drift (changes over time), and the formation of rater sub-communities. A key feature of the workshop is the comparison of NA results with those from traditional approaches such as the Rasch model.
NA is a flexible method that can handle nominal, dichotomous, ordinal, and numeric data. Unlike traditional models that rely on strong assumptions (e.g., local independence or unidimensionality), NA operates with minimal requirements, making it especially suitable for complex or non-standard rating contexts. Visualizations further enhance interpretability.
The workshop emphasizes hands-on experience using open-source R code and real datasets from published studies. A brief theoretical overview will also be provided. Participants are encouraged to bring their laptops and follow along.
Reading list:
Lamprianou (2018) in Educational and Psychological Measurement, 78(3), 430–459.
Lamprianou (2023) in Sociological Methods and Research, 55(1), 525–553.
Lamprianou (2025, in press) in Research Methods in Applied Linguistics.
Lamprianou et al. (2023) in Assessing Writing, 56, 100713.

Keynote Speech

9:30 - 10:15

All the World Is a Stage – Enter the Androids: Navigating Current and Future Challenges of International Baccalaureate Assessment

M. Glanville[1]

[1]International Baccalaureate, United Kingdom

The International Baccalaureate (IB) has an unusual role in aiming to offer an educational philosophy and approach for the whole world. In assessment terms, this presents the IB with a number of challenges – some of which will be familiar to national assessment systems and some which are unique. After providing a brief overview of the IB and its qualification I will reflect on these unusual and interesting matters.
However, the assessment landscape is not standing still and developments in technology and society are impacting on many of our foundations of education. In the second half of my presentation, I will talk about how the IB is thinking about these changes (AI being key amongst them) and speculate on what the known unknowns may be for future assessment.

Assessment Cultures I

12:00 - 12:30

How well are they? Assessing wellbeing.

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

[1]independent, United Kingdom
[2]Institute of Education, UCL, United Kingdom

Wellbeing is increasingly seen as an objective of education. But how can it be assessed? And should it be assessed? The presentation offers five conclusions for discussion. It starts by referring to the two main models of wellbeing - the “eudaimonic” model, related to what is judged to be worthwhile, and the “hedonic” model, with wellbeing denoting a balance of pleasurable feelings over unpleasant ones. It concludes (1) that the eudaimonic element is essential and cannot be eliminated. Wellbeing is presented as an “essentially contested concept”, but the presentation concludes (2) that assessment needs to accommodate important concepts like wellbeing, despite the absence of agreement on the definition of the construct. Turning to how wellbeing can be assessed, it argues (3) that understanding wellbeing and being able to discuss it with others are good – and can be tested – but are not the same as being well. It contends (4) that the subject is not the only source of evidence about how well he or she is. Finally, it asks whether we should assess wellbeing, concluding (5) that surveying collective wellbeing can serve the public good; but there are ethical dangers in drawing conclusions leading to action on individuals.

Private judgments, public stakes: How can reliable final assessment be achieved in a culture of autonomy?

L.V. Sandvik[1], M.B. Johansen[1], A.H. Kvistad[1], B. Svendsen[1]

[1]NTNU, Norway

In decentralized education systems, professional autonomy is often seen as a cornerstone of high-quality teaching. Yet in contexts where teachers are solely responsible for assigning final grades—such as Norway's standpunktvurdering—this autonomy can undermine reliability and fairness in summative assessment.

This paper draws on a national mixed-method study investigating the extent and nature of collaborative assessment practices, and how tensions between autonomy and reliability manifest in this final grading practices. The data consist of survey responses from 792 teachers and 34 focus group interviews across eight schools.
Findings show that assessment remains largely an individual practice. Teachers value autonomy and express trust in colleagues but rarely calibrate or align their summative judgments. Professional learning communities (PLCs) exist in name, yet lack the time, leadership, and shared purpose needed to impact assessment quality. Leadership plays a limited role in enabling collective practices.

We argue that valid and transparent assessment requires more than professional discretion; it depends on shared standards and structured dialogue. The paper calls for a cultural shift—from assessment as a private responsibility to a collective professional task—and contributes to international debates on how to strengthen assessment reliability in autonomous school systems.

Critical consciousness and well-being: Questioning and transforming teachers' relationship with assessment.

C. F. Correia[1]

[1]University College London - IoE, United Kingdom

This paper presents a conceptual analysis of the potential of critical consciousness to question and transform teachers' relationship with assessment.

The first part discusses impacts of performativity/managerialism on the construction of teachers' assessment identity and the adoption of certain assessment routines. The discussion will focus on tensions arising from teachers inhabiting multiple identities as educators, assessors, and carers. It will be argued that performativity/managerialism are tools for teacher oppression. They can promote internalised behaviours that support surveillance and distort the synergy between formative and summative assessment, in the classroom. These behaviours lead the reproduction of well-established assessment routines that can have significant costs for both teacher and student well-being.

The second part discusses the role of critical consciousness in promoting spaces for reflection and resistance in teacher-education/professional-development. Critical consciousness is a vehicle to reflect/deconstruct the rationales and assumptions behind well accepted assessment routines in the classroom. This reflection/deconstruction can be achieved through critical analysis of those routines, and the construction of a collective teacher assessor identity. These, in turn, support the development of teachers' political self-efficacy and promote critical action towards more healthy assessment practices that prioritise learning and well-being.

Assessment of Practical Skills I

12:00 - 12:30

Designing a Test to Strengthen the Functional Literacy of Students at Nazarbayev Intellectual Schools.

D. Sartauova[1], F. Khamza[1]

[1]Centre for Pedagogical Measurements under AEO Nazarbayev Intellectual Schools, Kazakhstan

This study explores the development and implementation of a digital assessment tool designed to enhance the functional literacy of students at Nazarbayev Intellectual Schools (NIS). The tool aims to foster self-directed learning and improve key literacy domains—Reading, Mathematical, and Scientific Literacy—through personalized, adaptive learning opportunities. The study draws on theories of self-directed learning (Knowles, 1975; Zimmerman, 2002) and modern test methodologies, including Item Response Theory (IRT) and Computerized Adaptive Testing (CAT), to ensure the tool's reliability and scalability. Aligned with international educational standards such as the OECD's PISA and UNESCO's definitions of literacy, the tool provides immediate feedback to help students identify and address learning gaps independently. The research uses a mixed-methods design, incorporating pre- and post-tests, surveys, and interviews to assess the tool's effectiveness. Expected outcomes include improvements in students' functional literacy skills, increased engagement, and validation of an innovative digital assessment model. The study also offers policy and pedagogical recommendations for scaling the tool across various educational systems, contributing to global discussions on digital literacy assessment and functional literacy development. Ultimately, the findings aim to inform educational policies in Europe and Central Asia, promoting lifelong learning and critical thinking skills.

The validity of virtual labs for assessing science practical skills

E. Walland[1], A. Rodrigues[2]

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

Various methods to assess secondary school science practical skills exist, and each has a different impact on teaching and learning. Direct assessment of practical skills through physical labs is not always possible for schools. Alternative to practical exams, where students do written papers instead, are an option. However, such written exams do not require students to manipulate equipment and there is potential for negative washback.

Our study uses a critical literature review to explore virtual labs as a potential means to assess science practical skills. Virtual labs are a popular teaching tool, however, in assessment they have received less attention. We focused on exploring the validity of virtual labs for the summative assessment of science practical skills. Using the Crooks et al. (1996) chain model as a theoretical framework, we elucidate threats, benefits and opportunities for the validity of virtual labs in summative assessment, compared with physical labs and alternative to practical exams.

We argue that virtual labs can potentially assess a subset of science practical skills, and because they require virtual manipulation of equipment, they represent an improvement over written examination. However, there are several threats to validity that need careful consideration, including the impact on classroom practices.

How are oracy skills assessed in England and internationally?

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

[1]Ofqual, United Kingdom

Effective speaking and listening skills are fundamental to fostering meaningful communication, collaboration, and understanding in society. Yet, oracy remains a complex construct to define, teach, and assess. Across many countries, oracy skills are integrated into curricula, but summative assessment practices vary significantly. These variations potentially influence the emphasis placed on oracy and the degree to which students develop these skills.

This study aimed to understand the different approaches to assessing oracy skills internationally. We reviewed the literature and a sample of qualifications to examine how oracy is defined, assessed, graded, and importantly controls put in place to support reliability and validity. We found that definitions of oracy varied significantly within the literature. Qualification purpose and the stakes of the test largely influenced the construct assessed, assessment methods, approaches to grading and quality assurance practices. Common modes of assessment included oral presentations and group discussions; however even when qualification purpose was similar, task design and duration varied.

We discuss key considerations for assessing oracy skills validly and reliably, drawing on international practices, and emphasize the critical role of purpose when designing oracy assessments. Implications for diverse student groups, including learners with speech and language difficulties are also addressed.

Psychometrics and Test Development I

12:00 - 12:30

How does different linkage plans affect the equating transformation?

M. Wiberg[1]

[1]Umeå University, Sweden

Test score equating is used to make score scales from different test forms comparable. It is common to use the nonequivalent group with anchor test design when equating test scores. The overall aim is to examine the equated values and the standard errors when different linkage plans have been used under different circumstances. We used test forms from the Swedish Scholastic Aptitude Test, which is a multiple-choice binary scored test used for college admissions. We also used simulations to examine different linkage plans under different conditions such as different mean and standard deviation among the test taker groups as well as if the average item difficulty varied on the different test forms. Preliminary results show that the equated values vary depending on which linkage plan is used and how the test forms varied. Practical implications and recommendations for how to choose a linkage plan in standardized achievement tests are given.

A Peculiarity in Educational Testing Practices

M. White[1]

[1]University of Oslo, Norway

This paper discusses a peculiarity in the institutionalized practices of educational testing. Namely, there is an inherent contradiction between guidelines for test development and test analysis. Tests are built using test blueprints that assign quotas based on characteristics of test items (e.g., sub-domain, open response, cognitive demand). This is because different types of items are assumed to contain unique, construct-relevant variation and this variation must be balanced when measuring the intended construct (i.e., a test of math ability must include some geometry items because those items measure a facet of math ability that is not captured by algebra items). Reflective measurement models, such as item response theory, on the other hand, assume that no individual item or subset of items contains unique, construct-relevant variation (i.e., the local independence assumption). The two sets of practices, therefore, make fundamentally contradictory assumptions. The implication is that reflective models do not properly separate construct-relevant variation from error, which is a necessary precondition for providing information about test reliability (i.e., the percentage of construct-relevant variation) and test validity. I argue for the need to align test development guidelines and the analysis of data from tests.

Effects of using IRT scoring on admission outcomes

P. Lyrén[1], I. Laukaityte[1], M. Wickberg[2], T. Lindquist[2], H. Pettersson[2]

[1]Umeå University, Sweden
[2]Swedish Council for Higher Education, Sweden

The SweSAT is used, alongside school grades, for selection to higher education in Sweden. Classical test theory (CTT) has been the main framework for the SweSAT development, but discussions about digitalization have highlighted the potential of using item response theory (IRT) as the main framework. In this study, we investigated the effects of using IRT-based SweSAT scores on admissions outcomes.
A simulated admission was used where IRT-based scores replaced the baseline CTT-based scores, and we examined overall effects as well effects on specific programs. The participants comprised 402,162 applicants and 797,227 test-takers.
In the simulation, 27% of applicants with SweSAT scores earned a higher score, 30% a lower score, and 43% remained unaffected. There is a reshuffling of 2,017 individuals between baseline and simulation, with 1,036 losing their place and 981 gaining a place. The impact on program composition is minimal, with slight changes in gender and age distribution. At the individual level, the proportion of those admitted in the baseline that were not admitted in the simulation varied between programs, from 0.4% in preschool teaching to 6.2% in dentistry.
In conclusion, using IRT-based scores has minimal impact on overall admissions outcomes and program composition.

e-Assessment I

12:00 - 12:30

Are schools ready for digital exams? – Creating a framework to evaluate digital readiness

S. Mistry[1], S. Kemmery[1]

[1]Cambridge University Press and Assessment, United Kingdom

This study investigated schools' readiness for delivering and sitting digital exams across international and UK markets by using an internally developed readiness framework.
The primary objectives were to understand the readiness landscape for digital exam delivery, capturing the use of digital in teaching and learning, and creating a framework for ongoing market analysis.
Employing a mixed-methods approach, the research included an online survey and in-depth interviews. Focus areas for development of the framework included hardware, connectivity, resources, student readiness, digital teaching, readiness timescale, risks, and willingness. Questions related to each area was scored from 1 to 4, with willingness to offer digital exams scored separately.
Results showed general acceptance of digital assessments, with significant regional differences in readiness and willingness. Only 22% of schools achieved a high readiness score, indicating considerable variance. Barriers for digital exam readiness and adoption included financial investment, efficacy concerns, resource implications and complexity. The session will share the regional data obtained across the digital exam readiness focus areas.
The study concluded that for delivery of digital exams by Cambridge, a phased digital exam adoption is necessary, requiring extensive training and support to address practical and attitudinal barriers.

Assessment mode effects and their relationship with item characteristics

C. Vidal Rodeiro[1], C.H. Lim[1]

[1]Cambridge University Press & Assessment, United Kingdom

Digital exams are becoming part of the future assessment landscape. They are being gradually rolled out and are mostly delivered alongside paper versions. However, offering exams simultaneously in both modes raises comparability concerns: will the same items in different modes assess the same constructs and function similarly? Ensuring students are not disadvantaged by assessment mode and everyone benefits equally from technological innovation is important.

This research investigated Differential Item Functioning (DIF) by mode (i.e., mode effects), using Partial Credit Models and data from Cambridge University Press & Assessment's digital mocks and paper tests in subjects including Computer Science and English Language. Items in both modes were also reviewed and classified according to their characteristics to investigate whether certain types were more likely to drive mode effects.

Our findings suggest that mode effects exist but are not extensive. Additionally, whether items were harder on paper or on-screen was not consistent within or across subjects, indicating that students are unlikely to be systematically disadvantaged due to assessment mode. On item characteristics we found, amongst other things, that reading items requiring "accessing and retrieving" information from a text were about twice as likely to exhibit DIF compared to items involving other text interactions.

HyPES project - assessment of learning in hybrid learning environments in higher education: what methods and tools?

N. Natasha[1], N. Younès[2], J. Pironom[2], R. Burton[1], L. Massou[3]

[1]University of Luxembourg, Luxembourg
[2]Université Clermont-Auvergne, France
[3]Université de Lorraine, France

This research was conducted as part of the HyPES project, which aims to study hybrid systems in higher education.
The study primarily builds on the results of a previous European project that developed a typology of hybrid systems based on five main dimensions: (1) integration of in-person and remote learning, (2) mediatization, (3) mediation, (4) the degree of openness of the system and (5) human support.
This earlier theoretical framework has been enhanced by research aimed at developing an integrative model to address the complex dynamics involved in co-constructing hybrid learning environments. Two significant modifications have been made: the inclusion of variables that capture the interaction between students and digital teaching and learning environments, and the addition of a dimension related to the assessment of learning.
The focus of this communication is on the study of this latter dimension. The responses of over 170 French-speaking higher education teachers to an online questionnaire were analyzed. The way teachers design assessments in hybrid systems and the digital tools they use for this purpose will be described.
As this research progresses, the typology of hybrid systems will be updated to incorporate this evaluative dimension, thereby characterizing the systems and assessing their quality.

National Tests and Examinations I

12:00 - 12:30

Beyond Grammar – A Holistic Approach to KS2 Writing Assessment

R. Clarkson[1]

[1]Anglia Ruskin University, United Kingdom

The current assessment of writing at the end of Key Stage 2 (KS2) in England places significant emphasis on spelling, punctuation, and grammar, yet concerns have been raised about whether this narrow focus adequately captures the full scope of writing proficiency. This paper critically examines the extent to which a grammar focussed assessment reflects students' broader writing abilities. This paper reviews recent research on writing assessment and explores alternative approaches. Within the context of a national curriculum review, practical recommendations are offered for adapting KS2 writing assessments to better reflect the complexity of real-world writing, ensuring that young learners develop not only grammatical accuracy but also the expressive and critical thinking skills needed for lifelong literacy. This paper argues for a shift towards a more balanced evaluation framework that values not just technical accuracy but also the richness of expression and communication in student writing.

Using Aberrant Responses to Identify Backwash in National Mathematics Tests

M. Mikite[1], G. Burgmanis[2]

[1]University of Latvia, Latvia
[2]LU Starpnozaru izglītības inovāciju centrs, Latvia

National tests are designed to assess students' mastery of curriculum standards, yet they can also shape instructional practices—a phenomenon known as backwash. While backwash can have both positive and negative effects, this study focuses on its potentially detrimental impact, where teaching is narrowly tailored to test formats, limiting deeper learning. To explore this, we apply aberrant response analysis to national mathematics exam data from two consecutive cohorts in Latvia. Aberrant responses—unexpected answers relative to a student's overall performance—can signal instructional or motivational inconsistencies. Using Rasch modelling and expert-coded item characteristics, we developed three person-fit indices: Instability, Understanding, and Drill. These indices were aggregated at the school level and compared across school types. Additionally, year-to-year correlations were used to examine the stability of response patterns. The findings indicate systematic differences across schools, with some exhibiting consistent patterns suggestive of overemphasis on procedural practice and underperformance on conceptual tasks. These results suggest that aberrant response analysis can serve as a diagnostic tool for identifying unintended instructional effects in national testing contexts, contributing to more informed and targeted educational improvements.

Artificial Intelligence and Assessment I

12:00 - 12:30

Principles of AI use in high stakes marking

J. Williamson[1]

[1]Ofqual, United Kingdom

This research explores the principles of using AI in marking for high stakes qualifications. AI systems may be used as the primary mechanism for determining a mark, or as part of quality assurance processes supporting human markers. Understanding the implications of AI use is critical for both policy and practice.

AI technologies continue to develop apace, and it is unknown how capabilities will change in the near or medium-term future. The logic of this research was to support clear thinking about uses of AI in assessment by focusing on core assessment principles. As such, the approach taken was to unpack the role of AI-supported marking within the framework of assessment validity arguments.

This presentation focuses specifically on the relationship between human academic judgement, explainability, transparency and assessment validity in the context of high stakes qualifications. It explores how this relationship varies with item type and assessment construct; the combination of AI and human marking; and the type of AI deployed (in particular, contrasting AI systems built upon foundation models like LLMs with earlier machine learning approaches). The research contributes a clearer understanding of the logic of various AI implementations, and the empirical evidence necessary to demonstrate their validity in practice.

Responsible Use of AI in Research and Assessment: Emphasizing Reliability through Repeated Evaluations

E. Papanastasiou[1], A. Stylianou Georgiou[1]

[1]University of Nicosia, Cyprus

The growing popularity of large language models (LLMs) in education has spurred their use in a range of tasks, including creating assessments and evaluating student responses. However, due to the potential for AI to generate misleading outputs, the European Commission has proposed guidelines emphasizing responsible usage of generative AI in research, focusing on reliability, honesty, respect, and accountability. These guidelines call for verifying and reproducing information produced by AI, ensuring transparency regarding AI use, recognizing the technology's limitations, and taking overall responsibility for the outputs. Building on these principles, the present study aimed to demonstrate a process for applying them in research, centering on evaluating how accurately LLMs reproduce a qualitative data analysis conducted by human researchers and how consistent their outputs are across multiple LLMs. The results of this study demonstrate an example of how the guidelines for responsible AI use can be applied for evaluating the use of AI in research and assessment. Moreover, they demonstrate that LLMs are not always error free, and that human oversight of the process and the evaluation of its results are imperative.

Holistic Assessment

12:00 - 12:30

Debunking a dichotomy: An analysis of future skills in knowledge-rich qualifications

I. Suto[1], S. Nelson[1], J. Roberts[1], A. Sidhu[1], L. Spence[1]

[1]Cambridge International Education, United Kingdom

A holistic education should nurture skills that are essential for thriving in a fast-changing world, such as systems thinking, kind and wicked problem-solving, and metacognition. These ‘future’ skills are deeply intertwined with conceptual and factual knowledge and should not be assessed in isolation. Curricula that are structured around long-established subject disciplines within the sciences, humanities, and languages, and assessed via written examinations, are often considered ‘knowledge rich.’ This paper investigates how they also foster future skills.

We conducted a systematic analysis of international A level qualifications in English Language, Geography, Physics, and Psychology. To syllabuses, specimen examination papers, and mark schemes, we applied a coding framework based on Marzano and Kendall’s (2007) New Taxonomy of Educational Objectives, which encompasses problem-solving and metacognitive skills. We incorporated additional codes for systems thinking components from the relevant domain of an environmental sustainability framework.

The study revealed a broad and rich coverage of higher-order thinking skills, with variation across subjects. For example, although systems thinking is absent from formal assessment objectives, it is present in Geography, Physics, and Psychology examinations. The study also highlighted that general and subject-specific exam techniques inherently demand many higher order skills, further demonstrating the skills’ integration within these qualifications.

Understanding the construct of ‘future skills’: some critical reflections emerging from a comparison of recent and older future skills frameworks

F. Constantinou[1], A. Miranda[1], S. Pirola[1]

[1]Cambridge University Press & Assessment, United Kingdom

A key mission of education is to equip learners with ‘future skills’, namely, skills essential for navigating the future. To evaluate whether education has succeeded in achieving this mission, appropriate assessment instruments need to be developed. This paper seeks to inform the assessment of students’ preparedness for the future by sharing some critical reflections on the construct of future skills. The reflections emerged from a study which aimed to investigate whether any new future skills arose in recent years. The study, which took the form of a systematic review, identified 35 future skills frameworks developed during 2021-2024. It subsequently mapped the 562 skills constituting these frameworks to an existing meta-framework. The reference meta-framework resulted from the analysis of older frameworks (2002-2021) and, as such, served as a useful benchmark for assessing the novelty of the skills in this study. The mapping process led to several interesting observations about the construct of future skills. These observations were synthesised into a typology of attributes to help clarify the nature of the target construct. These attributes included: responsive, evolving, heterogeneous, engineered, atheoretical, open-ended, elusive and repetitive. This presentation will introduce and exemplify the typology and will also reflect on its implications for assessment.

Creativity at the Crossroads of STEM and Non-STEM Subjects: Pathways to Holistic Education

G. Sultanova[1], A. Zhapparova[1], Y. Almenov[1]

[1]Center for Pedagogical Measurements, NIS, Kazakhstan

Holistic development involves nurturing a student's intellectual, emotional, social, and creative skills, recognizing their interconnectedness and equal importance for personal growth. Creativity is central to this process, enabling students to generate innovative ideas, solve complex problems, and adapt to new challenges. In particular, Science, Technology, Engineering, and Mathematics (STEM) education benefits from a balanced approach that integrates diverse knowledge domains, as these offer complementary skills essential for fostering creative thinking. This study explores how academic achievement in four STEM subjects (Mathematics, Physics, Chemistry, Biology) and four non-STEM subjects (Kazakh, Russian, English, History) affects creativity in STEM secondary education, using PISA 2022 data from 2,435 students in Kazakhstan. Path analysis demonstrated direct effects of Mathematics, English, Biology, and History on creativity, with other subjects contributing indirectly. The model showed an excellent fit (??????????????????=0.010, ????????????????=0.004, ????????????=1.000, TLI=0.999), with creativity's variance (R^2=7%) reflecting its complex and multidimensional nature. The findings underscore the importance of balanced curricula integrating STEM and non-STEM subjects. As the first study of its kind in Kazakhstan's STEM school context, it provides evidence to inform educational policies that aim to promote creativity through a more holistic lens in secondary education.

Discussion Group 1

14:30 - 15:30

How do we reconfigure assessment to meet the challenges of cultures driven by technology?

E. Andressen[1], G. Clark[2], S. Shaw[3, 4]

[1]Trinity College, United Kingdom
[2]SQA, United Kingdom
[3]Institute of Education, University College London, United Kingdom
[4]CIEA, United Kingdom

Whilst the fourth, information technology-led, industrial revolution brought with it the concept of a knowledge economy, it also brought the means for knowledge to be stored and made available more easily than at any time in the past. This has had a significant impact on many aspects of society and cultures. As a consequence, the current landscape of education is undergoing a profound transformation, driven by rapid technological advancements and an increased focus on the needs of the individual and how learning and assessment must adapt to meet those demands. As we navigate cultural change, it is essential to re-examine our approaches to assessment and consider how it can be reconfigured to meet the demands of this increasingly technology-driven world. This discussion group will explore the implications of technological advancements on the purposes of and assessment practices in national examinations, including the potential of artificial intelligence and data to enhance or disrupt traditional approaches. By sharing perspectives and experiences, participants will contribute to a deeper understanding of the complex issues relating to the interface between cultures and assessment, driven by technology.

Discussion Group 2

14:30 - 15:30

The ‘Assessment Cultures’ SIG – who are we and what do we want to be talking about?

C. Schneider[1], S. Passeport[2, 3], M.S. Syverud[4], A. Watts[5], N. Younès[6]

[1]University of Trier, Germany, Germany
[2]University of Dundee, United Kingdom
[3]Tilburg University, Netherlands
[4]University of South-Eastern Norway (USN), Norway
[5]University of Cambridge, UK, United Kingdom
[6]Université Clermont-Auvergne, France

This discussion group targets at AEA-E-members who have signed up for the ‘Assessment Cultures’ SIG, those considering to do so and everybody sharing the an interest in what ‘Assessment Cultures’ are, how they differ across jurisdictions and educational contexts, and how they translate into everyday assessment practises.

The DG aims at contributing to transform the SIG into a living community in which its members would actively network with each other. It strives to identify interests of SIG members and encourage or intensify discourse amongst members on topics of interest.

The session will be informed by data gathered in a quantitative survey of SIG members administered timely before the conference. In a 10-minute introductory presentation, survey results will be presented. Based thereon, attendants will be encouraged to discuss in focus groups on topics of interest. Following an “unconference” approach, the focus groups themselves should ideally identify discussion questions in a bottom-up fashion. Given the limited DG timeframe, however, the authors will prepare some impulse questions for a number of foci. The DG will conclude with a plenary session in which focus groups will report on their exchange and, potentially, arrange on how to bring their discussion forward in future SIG activities.

Discussion Group 3

14:30 - 15:30

AI-supported translation tool for linguistically inclusive classrooms

T. Burner[1]

[1]University of South-Eastern Norway, Norway

In this research study, an AI-supported translation tool is tested out with up to 300 students at several schools. The tool is part of a GDPR safe AI-powered platform built on AI-models like GPT (Generative Pre-Training Transformer). A custom version of Azure OpenAI is used to interpret images and other document formats. The AI can answer questions from students and provide immediate feedback on any text, audio or multimodal file. The translation tool enables immediate translation of a text or parts of a text to any language. The study posed the questions whether, to what extent and how students use the translation tool. Data are collected in two ways: screen tracking and recording of all the student iPads/PCs during lessons, and focus group interviews with purposefully selected samples of students.

The session starts with a short presentation of the study, followed by discussion and reflection questions regarding the use of AI translation tool to support students in linguistically and culturally diverse classrooms. The ultimate goal of the tool is to include more students through increased understanding and involvement with texts provided by the teacher and by peers.

Discussion Group 4

14:30 - 15:30

Should we have confidence in high-stakes assessments?

I. Nisbet[1], M. Richardson[2], S. Shaw[3, 4], L. Wiseman[5]

[1]Educational Research Consultant, United Kingdom
[2]UCL Institute of Education, United Kingdom
[3]Institute of Education, University College London, United Kingdom
[4]Chartered Institute of Educational Assessors, United Kingdom
[5]Research Fellow University of Glasgow, United Kingdom

Most societies with a compulsory education system depend on high-stakes assessments as a means for certification and as a recognised part of employment, educational and social structures. In order to make good selections for employment or other opportunities and to trust and have confidence in the fitness to practise of ourselves and others we need a recognisable/accepted system of accreditation (Billington, 2011).

This discussion group is an opportunity to consider perceptions of confidence in high-stakes assessments of any kind and in any country or jurisdiction. We will include discussion of the extent to which participants (in the room) and other stakeholders have such confidence; the factors contributing to it; the evidence supporting it; and whether there is a need for it.

Participants will engage in small-group, round table discussions exploring the factors contributing to perceptions of confidence and trust. They will be invited to consider questions: What gives you confidence in high-stakes assessment? Does anything undermine your confidence in high-stakes assessment? A plenary session will provide opportunities for all to share views, comments and questions.

Billington, L. (2011) Public Trust and High Stakes Assessment. Available at: https://filestore.aqa.org.uk/content/research/CERP-RP-LB-05112007_0.pdf (accessed 27 March 2025)

Discussion Group 5

14:30 - 15:30

Integrating assessment and curriculum design: Promoting authentic connections between learning and assessment

J. Hayman[1], V. Westwell[1], J. Scanlon[1]

[1]International Baccalaureate, United Kingdom

In education it is important to link teaching and assessment to foster higher order thinking skills (HOTS). One international program incorporates both ‘in-school’ assessments and optional external e-Assessments consisting of ePortfolio work and on-screen examinations, with the aim of developing and assessing critical thinking.

Curriculum and assessment design must account for delivery in diverse school settings. While e-Assessments are not all mandatory, they reflect the standards expected at the end of the five-year course and influence curriculum design. On-screen examinations are specifically crafted to assess higher order thinking skills in a digital environment and are developed collaboratively by subject managers and multimedia developers. All e-Assessments serve as models for schools, whether they register candidates for them or not.

To ensure consistency and clarity, assessment criteria must be flexible enough to apply to both internal and external assessments. A panel discussion will explore how connections between teaching and assessment have been achieved in on-screen examinations using examples from science, mathematics, and interdisciplinary subjects.
The panel will discuss the importance of fostering higher order thinking skills and the backwash effect of external assessments on curriculum design, highlighting how these assessments can be part of the learning process.

Discussion Group 6

14:30 - 15:30

Beyond Alignment: Assessment at the Intersection of Constructiveness, Accountability, and Tacit–Explicit Knowledge Dynamics

Y.P. Hsiao[1], A. Kramer[1], M. Stienstra[1]

[1]Tilburg University, Netherlands

This discussion focuses on assessment policy input regarding three propositions for integrating constructiveness, accountability, and tacit–explicit knowledge dynamics into assessment design at both program and course levels. Constructive alignment, a widely used framework, aligns learning outcomes, instructional activities, and assessment tasks, promoting active student engagement with both explicit knowledge (formal, codified learning) and tacit knowledge (informal, experiential understanding). While alignment is crucial for assessment validity, the "constructive" element remains conceptual in practice, particularly in assessment design. The discussion stresses the need to intentionally integrate both alignment and constructiveness to enhance curriculum development and deepen student engagement. Moreover, higher education has traditionally emphasized explicit knowledge transfer, often neglecting tacit knowledge and the development of students' evaluative skills. Participants are invited to explore perspectives from various stakeholders to discuss how self-regulatory practices, such as self-assessment and self-feedback, can help students navigate complex, real-world environments beyond formal education. Tacit knowledge is central to criterion knowledge—the ability to make informed evaluative judgments. Assessment designs should encourage the integration of both tacit and explicit knowledge. The discussion invites participants to experience holistic assessment, which evaluates quality interactively by blending these knowledge types, is seen as a promising approach to fostering adaptive expertise and meaningful learning.

Discussion Group 7

14:30 - 15:30

Artificial Intelligence for Holistic Learning: Challenges and Possibilities in Assessing Collaboration and Empathy

G. Sultanova[1], I. Suto[2], C. Hutchinson[3]

[1]Center for Pedagogical Measurements, NIS, Kazakhstan
[2]Cambridge International Education, United Kingdom
[3]University of Glasgow, United Kingdom

Artificial Intelligence (AI) is increasingly being used in education to plan and support teaching, personalise learning, automate feedback, and measure learning outcomes. However, its potential to support holistic assessment, which evaluates not only traditional ‘academic’ knowledge and skills, but also social and emotional skills, remains largely unexplored.
This discussion group will delve into AI's potential for assessing learning and development in this broader range of skills. In small interactive groups, participants will reflect on the challenges and opportunities of using AI for assessing collaboration and empathy, with an emphasis on practical and ethical considerations.
The session will consider what learning and progression in these skills might look like in educational settings from nursery through to colleges and the workplace, and whether and how AI can be integrated into existing assessment practices and contribute to ethical, human-centred approaches across these contexts.
Discussion will focus on the benefits, limitations and risks of assessment using AI, as well as the ethical, pedagogical and technical challenges, especially its possible impact on learners at different stages of their education. Participants will leave with practical insights and principles about responsibly integrating AI into holistic assessment practices, where they consider this approach might be practical and ethical.

Discussion Group 8

14:30 - 15:30

Universal Design for Assessment: What it is, what it is not and the tensions it must balance

K. Parasuram[1], M. Manzo[2]

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

Universal design for learning (UDL) is a key reference point for educational authorities worldwide, with its principles now widely adopted. However, although progress has been made in applying universal design principles in the classroom, a noticeable gap remains in their application to formative and summative assessment. The goal of Universal design for assessment (UDA) is to ensure all students can fully engage with and participate in assessments, demonstrating their knowledge, skills, and abilities without unnecessary hindrances. UDA encourages the use of multiple means of engagement, representation, and expression in assessment design to support diverse learners in demonstrating their learning, while still ensuring assessments remain valid. UDA promotes identifying gaps in ancillary skills to inform learning, however in summative contexts, the impact of ancillary skills must be minimised to ensure fairness.

The discussion group will explore practical strategies for incorporating flexible methods and cultural inclusivity into assessments while maintaining rigour. Participants will reflect on the balance between accommodating learning needs, flexible methods, cultural diversity and inclusivity in assessment design against assessment validity, comparability, standardisation and manageability. The discussion group will give participants practical insights to implement universal design for assessment both in international as well as in their own contexts.

Discussion Group 9

14:30 - 15:30

From grades to growth: Reframing the dialogue on learning, assessment and mastery

K. Blichfeldt[1], E.W. Hartberg[1], I. Jacobsen[1], K. Haaland[1], T.S. Wille[2]

[1]Faculty of Education, Inland Norway University, Norway
[2]The Education Agency, Oslo Municipality, Norway

This discussion group builds on findings from the Norwegian DOLA study (Dialogues on Learning and Assessment) and early insights from the DRIVE project. It explores how current assessment practices, particularly in high-stakes, summative-oriented contexts, can undermine the goals of education: meaningful learning, motivation, and student agency.
While formative assessment is emphasized in policy, DOLA1 and DOLA2 show that both early career and experienced teachers struggle to integrate it within systems that prioritize grades. Students often view formative feedback as a disguised grade rather than a tool for growth. These findings echo global concerns about how high-stakes testing distorts assessment practice, replacing genuine learning dialogue with procedural routines.
The session invites researchers, teachers, and policymakers to rethink assessment in ways that promote learning. Key themes include bridging the gap between curriculum and practice, addressing teacher tensions, reimagining student-teacher assessment dialogue, and challenging an assessment culture shaped by summative logics.
Participants will engage in discussions, reflect on shared dilemmas, explore real-world cases, and co-develop ideas for assessment that support motivation and mastery. The session aligns with AEA's call to design future assessment practices that are student-centered, learning-driven, and developmentally meaningful.

Discussion Group 10

14:30 - 15:30

Why should assessment of vocational or professional education and training matter to this conference?

L. Gray[1], R. Conway[2], N. Mellor[3], H. Debowski[4], V. Morgan[5]

[1]Independent consultant, United Kingdom
[2]NCFE, United Kingdom
[3]Chartered Insurance Institute, United Kingdom
[4]Central Examinations Board, Poland
[5]City and Guilds, United Kingdom

The diverse nature of assessments in vocational education and training, which spans various educational levels and industries, creates challenges in establishing common standards across Europe, and can hinder professional discussion of assessment issues. Full and informed debate about the nature of educational assessment in this field may be hampered by this lack of common understandings. This discussion group will explore key issues such as the unique challenges of assessment in vocational education, the impact of assessment design, and the interconnected nature of learning programmes and their assessments in this educational sector. The discussion will use a goldfish bowl format to facilitate dynamic conversations and share knowledge about evaluating the validity, reliability, and fitness for purpose of VET assessments. Key questions to be discussed will include why VET assessment should matter to the conference and what makes these assessments unique? Participants will also have an opportunity to contribute to developing a common language and suggest focus areas for the newly established Special Interest Group (SiG) on 'Assessment for Qualifications for Work.' Potential focus areas include differences in assessment approaches for various learner groups and the emphasis on transferable skills for changing employment landscapes.

International Assessments I

15:30 - 16:00

Standards in the wild: Setting and maintaining standards in a high-stakes global context

L. Miller[1], S. Hughes[1]

[1]Pearson, United Kingdom

Maintaining standards in a high-stakes global English language testing context presents unique challenges. Testing organisations must meet the needs of multiple stakeholders for diverse testing populations around the world. Further, multiple tests operate within this global context, each offering its own evidence of standards but without a centralised point of oversight to ensure consistency or comparability across tests. This context demands robust processes to ensure test score users are provided with sufficient evidence to trust the standard of test scores. This presentation outlines the steps taken to establish and maintain standards for a high-stakes global English language test, while supporting test score users in understanding and operationalising these standards.

This process requires triangulating multiple research streams including alignment research that considers how test scores relate to language proficiency frameworks, score concordance research that uses statistical analysis to relate test scores from one test to test scores on a similarly purposed test, and stakeholder engagement research that involves structured dialogue to explore the needs and expectations of score users alongside detailed exemplification of performance standards.

In building and maintaining trust through transparent and rigorous standard setting processes, we can ensure that the standards remain credible and respected across diverse educational landscapes.

Assessment Cultures II

15:30 - 16:00

Determinants and consequences of behavioral grades in schools

R. Alne[1], A. Fidjeland[1, 2], R.B. Reiling[3]

[1]Nordic Institute for Studies of Innovation, Research and Innovation, Norway
[2]University of Stavanger, Norway
[3]Statistics Norway, Norway

We use merged administrative and registry data from two large counties in Norway to examine how formalized behavioral assessments influence teacher-assigned grades in high schools. By comparing teacher-assigned grades to anonymously graded external exams, we exploit an institutional practice that evaluates behavior (orderliness and conduct) separately from academics. Our results show that receiving a lowered behavioral grade significantly reduces students' final in-class evaluations, even after controlling for exam-measured skill. Moreover, the penalty is larger for orderliness than for conduct, suggesting that teachers weigh organizational skills and punctuality more heavily than interpersonal behavior when assigning academic grades. Although boys and immigrant-background students are more likely to receive reduced behavioral grades, the size of the penalty remains comparable across these subgroups. Overall, our findings raise concerns about the fairness of teacher assessments and underscore the challenges of fully isolating academic performance from behavioral factors.

Artificial Intelligence and Assessment II

15:30 - 16:00

A Comparative Analysis of AI-Assisted and Human Interpretations of the Bloom's Taxonomy in the Classification of Test Items

O. Adetoro[1], D.A. Lawal[1]

[1]The West African Examinations Council, Nigeria

The use of artificial intelligence AI in educational assessment is redefining the assessment landscape as many examiners now leverage on it to improve the assessment process. This study was a comparative analysis of AI-assisted and human interpretations of the Bloom's Taxonomy in the classification of test items. Three hypotheses were formulated to guide the study. Ex-post facto research design was adopted. Population of the study comprised all WASSCE 2024 test items in Physics, Chemistry and Biology. All the 50 multiple choice items in each of the three subjects in WASSCE 2024 were considered. Bloom's Taxonomy was interpreted in the classification of the test items. Seasoned moderators (five for Chemistry; six for each of Biology and Physics) were used for the item classification. Regression and ANOVA were used for data analysis. Results revealed a moderate relationship between AI-assisted and human classification of WASSCE test items in Biology. Also, a moderate relationship between AI and human classifications in Physics (r=0.525); and Chemistry (r=0.267). It was recommended among others that examiners should not solely rely on AI-assisted decisions.

Joining Forces: Teachers and AI in Early Reading Risk Detection

P.H. Uppstad[1], B. Walgermo[1], N. Foldnes[1], K. Bjønnes[2], A. Aarflot[2]

[1]University of Stavanger, Norway
[2]Municipality of Oslo, Norway

Early identification of students at risk for reading difficulties is essential for effective instruction. However, first-grade screenings conducted at the beginning of the school year often fail to identify more than 50% of students who will struggle with reading by year's end. Research shows that while teachers can identify some poor readers missed by screening tools, they also overlook others. This study explores the prediction of teachers' professional qualitative evaluations of students' risk of reading difficulties at two time points relative to two AI-based screening approaches administered during first grade: a gameplay-derived risk algorithm and an adaptive reading screening test. Regression analyses of risk status for 655 first grade students across four time points revealed that the AI gameplay algorithm in November added predictive value beyond teachers' September evaluations. However, by March, teachers' updated evaluations rendered the AI predictions redundant. These findings are discussed in the context of teacher knowledge development and intervention, suggesting that combining teacher evaluations and screening scores may offer a more reliable method for identifying reading risks. Teachers' perspectives on the AI algorithm, gathered through focus group interviews, are also discussed. The study is a collaboration between XXX University and XXX muncipality in Norway.

Comparative Judgement I

15:30 - 16:00

Simplified pairs comparative judgement method for equating via expert judgement of script quality – further evidence of accuracy and bias

M. Curcin[1], M.W. Lee[1]

[1]Ofqual, United Kingdom

Comparative judgement (CJ) methods have been explored for use in different contexts, including for standard maintaining/equating, i.e., mapping scores on one test to equivalent scores on another test via expert judgement of the quality of student work (scripts) from those tests. In this use, CJ methods assume that judges can compensate for differences in difficulty between test forms when judging relative script quality.

Despite high reliability and plausibility of CJ outcomes, evidence is limited regarding the extent of CJ accuracy in replicating true difficulty relationship between different test forms compared to traditional statistical equating. Furthermore, there is some evidence of bias in CJ outcomes, whereby judges may not be able to adequately compensate for differences in test difficulty, judging performances on more difficult tests more severely than those on easier tests.

In this presentation, we expand the evidence base about the accuracy and bias of CJ methods for standard maintaining, focusing on comparing the outcomes of the simplified pairs method, applied to test forms with known difficulty relationship established through common-item IRT equating, with the IRT outcomes. We also considered the impact on outcomes of different script sampling approaches, number of judgements collected, judges' level of expertise and paper structure.

e-Assessment II

15:30 - 16:00

Unpacking Student Performance: The influence of Process data on International English Language Exam Results

L. Liu[1], B. Ashworth[1], H. Dalton[1]

[1]Pearson, United Kingdom

The transition from traditional paper-based assessments to digital platforms provides opportunities to investigate students' test-taking strategies through process (log file) data. Although process data has been extensively studied in low-stakes, large-scale international assessments, significant gaps remain regarding its applicability in high-stakes contexts and among test-accommodated learners. This study addresses these gaps by examining the relationship between process variables (time allocation, editing duration, and item attempt status) and exam performance in the 2023 International GCSE English Language examination. Data included responses from 526 students (average age approximately 16-year-old) across six countries.

Descriptive analysis revealed notable subgroup differences: females outperformed males on higher-mark questions despite spending less time overall but more on editing; learners receiving accommodations performed lower on higher-mark items and spent less effort editing these items. Utilising Least Absolute Shrinkage and Selection Operator (LASSO) regression to cross validate the Multiple Linear Regression (MLR) findings, the analysis highlighted that productive interactions, particularly strategic editing and effective word count management, significantly predicted higher performance. Conversely, excessive time spent correlated negatively with performance, particularly among older and those who partially attempted items. This research contributes by providing a methodological framework applicable to high-stakes exams and guiding future research towards adopting similar analytic approaches.

Student approaches to annotation in onscreen assessments: Exploring annotation techniques and their impact on test-taker strategies

Z. Mair[1], I. Custodio[1]

[1]Pearson, United Kingdom

In the current educational landscape, and as part of a transition to more digital high-stakes assessments, it is important to understand test-taker strategies across onscreen and paper-based formats. This research explores students' annotation strategies on source texts in English Language and English Literature exams both onscreen and on paper. Using a multi-phase approach, the visual outputs and student thought processes when making annotations are compared across text formats. Phase one included students from two schools (School A: n = 74; School B: n = 79) and compared annotation strategies used in International GCSE English Literature and English Language examinations. Preliminary findings showed that the percentage of students using annotations was consistent across paper-based and onscreen formats. The annotation strategies (as might be expected) differed visually, but with seemingly similar underlying motivations. Phase two involved post-test, semi-structured interviews to understand students' motivations for annotating and their perceptions of the impact of onscreen annotation tools’ on accessing, interpreting, and interacting with test material. This session provides insights into the comparability of student experience, test-taker strategies, perceptions and motivations when interacting with texts onscreen or on paper, thereby contributing to the overall understanding of assessment fairness and validity across different formats.

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

15:30 - 16:00

Curriculum for Wales: Insights from the Inaugural Stakeholder Surveys

G. Parkinson[1], B. Tylden-Smith[1], H. Thomas[2], J. Davis[2], L. Verdasco Menendez[3], C. Sinnema[4]

[1]AlphaPlus, United Kingdom
[2]Arad Research, United Kingdom
[3]AQA, United Kingdom
[4]The University of Auckland, New Zealand

The Welsh Government introduced a new Curriculum for Wales (CfW) in September 2022. The intent of the CfW is that individual schools are empowered to develop their own curriculum and assessment that are inclusive for all, while still ensuring all learners progress in line with nationally defined requirements. Following the initial years of implementation of any major system change, it is prudent to evaluate its impact. Surveys were developed to source feedback from the key stakeholders: learners in Reception to Year 9, parents and carers, practitioners, and school senior leaders. The surveys and the analysis conducted on it comprise the first nationwide opportunity to gather and review feedback on the implementation of the new CfW. This presentation details the key design considerations, approaches to and limitations of the analysis, key findings, and potential areas of development for the second survey planned for 2026/27. More specifically, it will cover key findings on the aspects of the curriculum addressed by the survey that featured the strongest levels of agreement or disagreement, the major themes emerging from the open text responses, and details of which population subgroups consistently responded differently to the rest of the population.

What Shapes Realisation of the New Curriculum for Wales? Understanding the Perceptions of Teachers, Learners and Parents/Carers Using a New Theory of Change

L. Verdasco Menendez[1], B. Tylden-Smith[2], C. Sinnema[3, 4], H. Thomas[4], G. Parkinson[2], J. Davis[4]

[1]AQA, United Kingdom
[2]AlphaPlus, United Kingdom
[3]The University of Auckland, New Zealand
[4]Arad, United Kingdom

The Welsh Government has developed a new Curriculum for Wales (CfW), operationalised as a theory of change (ToC) that argues CfW can shape learner experience and outcomes, parents' perceptions and practitioners' methods for the long term. This study assessed public perception of CfW and its impact on educational standards, as described in the ToC.
Two surveys captured insights from a) senior leaders and practitioners (SLP) and b) Reception–Year 9 learners and their parents/carers. Questions were mapped to ToC themes, responses weighted, and themes evaluated using confirmatory factor analysis and structural equation modelling.
For SLP responses, three hypotheses tested whether the ToC predicted pedagogy changes, quality of pedagogy and assessment, and learner outcomes, respectively. For parents/carers, it was hypothesised that perceptions of progress were predicted by trust in the school and satisfaction with information provided, collaboration on curriculum design, and degree of parent agency. For learners, it was hypothesised that perceptions of progress rested on learner collaboration on curriculum design, trust in school, and wellbeing.
This presentation will discuss findings and consider whether the ToC is a useful framework for evaluating CfW's impact. It should be of interest for those undertaking evaluations in other settings, particularly other ToC-driven curriculum reforms.

Psychometrics and Test Development II

15:30 - 16:00

Assessing Beyond the Score: The Role of Differential Item Functioning in Valid and Fair English Language Testing

J. Topham[1]

[1]Cambridge University Press and Assessment, United Kingdom

Cambridge University Press and Assessment has had an ongoing programme of work to collate and build the body of evidence on the validity of a suite of English language tests. Part of this work has been developing an operational Differential Item Functioning (DIF) analysis to assess bias toward, or against, gender, age and first language groups. The aim of this presentation is to show how DIF analysis has been implemented in an operational context for English language tests, focussing on the methodologies used and the specific considerations required given the construct being measured. It will also summarise the processes followed once items have been identified as potentially disadvantaging candidates from a specific subgroup of the test taker population. Operationalising DIF analysis has provided a wealth of data, giving an opportunity to monitor for long-term trends in bias. Therefore, this presentation will also explore over-time comparisons and the relationship between DIF outcomes and particular features of an item, such as test part or item type. It will be argued that these relationships can highlight ways in which tests may be unfairly biasing against certain groups, that may not be obvious when assessing individual items in isolation.

Fairness & Social Justice I

16:45 - 17:15

Predictive Validity of Unified National Testing in Kazakhstan

E. Kardanova[1], L. Shinetova[2], B. Abdrassilov[2], A. Ivanova[1], D. Orlov[3]

[1]National Research University Higher School of Economics, Russia
[2]National Testing Center of the Ministry of Science and Higher Education, Kazakhstan
[3]National Testing Center, Kazakhstan

The research aims to investigate the predictive validity of the Unified National Testing (UNT), a national admission exam in Kazakhstan. Introduced in 2004, UNT is a mandatory exam for admission to universities in Kazakhstan. More than 200,000 graduates take the UNT every year. Despite its long history and importance, there has been no research into the predictive validity of this test.
Sample for this study consists of 29,855 students who took UNT with combination of math and physics exams. Such a combination of exams is required to enter higher education institutions with IT and engineering majors. The data includes (1) UNT students' scores (a total score for the exam as well as separate scores for all subjects), (2) their GPA for the first session at the institute, and separate grades for math and physics subjects, and (3) some contextual information.
The preliminary results show moderate correlations between UNT grades and academic performance of students after the first session at the university, which is consistent with the results of the predictive validity studies of other admission tests (e.g. SAT). The detailed results of the UNT predictive validity study across different student groups (gender, language of testing, etc.) will be presented.

Gender Gap in Test Scores: Do Different Guessing Approaches and Test Creators' Gender Have an Impact?

L. Firtova[1]

[1]Scio, Czech Republic

This study examines gender differences in test scores in the General Academic Prerequisites test, the most widely used university admission test in the Czech Republic. By analysing five test versions from the 2024/2025 academic year, we found that, on average, male test-takers outperformed female test-takers. One contributing factor appears to be differing guessing strategies. Research indicates that women are generally more risk-averse, making them less likely to guess—particularly when a penalty for incorrect answers is imposed (such as a −1/3 point deduction in our case). In our study, the average omission rate was 28.5% for female test-takers, compared to 24.3% for male test-takers. This difference remained significant even after controlling for overall ability. While removing guessing penalties might reduce gender gaps in test scores, it could also negatively affect test reliability. We also explored potential gender bias in test content. Defining gender-specific content presents methodological challenges, so we focused our analysis on exploring the impact of question writers' gender, assuming it may inherently influence test content. We found that on average, both male and female test takers achieved slightly higher scores on female-authored questions. However, the gender gap in test scores was more pronounced in male-authored questions.

Investigating the Validity and Fairness of the Mandarin Leaving Certificate Examination in Ireland

X. Huang[1,1], D. Murchan[1]

[1]Trinity College Dublin, Ireland

The introduction of Mandarin Chinese into Ireland's Leaving Certificate examination in 2022 marked a significant milestone in diversifying national language education. This mixed-methods study examines Ordinary and Higher Level written papers from 2022–2024 to evaluate the extent to which they provide valid, fair measurement and reflect curricular aims.Through qualitative document analysis every reading, listening and writing item was coded for format, targeted skill and theme. In addition, shifts in visual scaffolding, cultural references and phrasing were traced across years. Data triangulation using semi structured interviews with three experienced Mandarin teachers illuminated classroom impact and perceived equity. Findings indicate that Ordinary Level papers continue to rely on highly controlled tasks, while Higher Level papers demand broader production of language without commensurate scaffolding. Both tiers privilege form focused exercises, limiting authentic communicative language use and disadvantaging non heritage learners. A persistent gap emerges between the published curricular aspirations of communicative competence and actual examination practice. We propose clearer level specific progression criteria, more open ended communicative tasks, and targeted teacher guidance to ensure that the Mandarin Leaving Certificate curriculum and examination are more closely aligned, provide better supports for learning, and evaluate student performance in a fair manner.

Artificial Intelligence and Assessment III

16:45 - 17:15

Evaluating AI assisted auto-marking systems using a range of item types

M. Frazer[1], S. Mistry[2]

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

Background
Researchers from Cambridge University Press and Assessment conducted trials into the use of auto-marking systems to support examiners. Pilot studies were conducted with two commercial providers of AI assisted auto-marking technology to discover the effectiveness and potential utility of using their systems. Following the initial studies a second phase of exploration took place involving a wider range of item types.
In phase 2, Supplier 2 was provided with a set of digital data relating to 89 items taken from 9 papers. The data were digital responses to these items which had been entered by candidates participating in in the Cambridge Digital Mocks Service January-March 2025 series.

Findings
1) The pilot studies showed that using AI assisted auto-marking to support examiners in the marking of medium and longer tariff items shows initial promise.
2) Agreement between examiner awarded marks and auto-marker awarded marks was acceptable in the case of 4 out of 6 items.
3) In the first instance, auto-marking may be used to support the work of the examiners in the Cambridge Digital Mocks Service.

Artificial Intelligence-Driven Test Item Generation: Enhancing Quality and Efficiency in Assessment Design

R. Ekpo[1], R. Ekpo[1]

[1]The West African Examinations Council, Nigeria

Artificial intelligence (AI) has revolutionized test item generation, enhanced efficiency and quality while overcoming human bias and inefficiency in conventional methods. AI-driven approaches, particularly those using natural language processing (NLP) and machine learning (ML), ensure reliable, valid, and adaptable assessments. This study explores AI's role in interpreting learning objectives, cognitive-affective-psychomotor taxonomies, and curriculum standards to generate relevant test items. AI also enables real-time calibration, fault detection, and adaptive testing, improving assessment quality. However, ethical concern-such as transparency, fairness, and bias-remain debated. Through a theory synthesis approach, this paper integrates validity standards, generative AI, and Automated Item Generation (AIG), comparing traditional and AI-driven methods. A systematic review of peer-reviewed studies (2018-2024) highlights AI's potential in assessment innovation while emphasizing integrity and fairness. The study recommends human-AI collaboration and AI-driven analytics (e.g., predictive analytics) for early detection of flawed items, ensuring balanced and effective assessment practices.

Investigating the Validity of ChatGPT in the Assessment of Open-Ended University Tasks: A Comparative Study with Human Grading

G.C. Pillera[1]

[1]University of Catania, Italy

The recent advancements in Generative AI (GenAI) technologies have prompted policy-makers, scholars, and education professionals to reflect on their potential applications in the field of assessment: for monitoring learning, designing assessment tools, supporting diagnostic, predictive, formative, and summative evaluation. In this fast-evolving landscape – rich in opportunities yet not devoid of risks – the study presented investigates the concurrent validity of ChatGPT in assessing academic open-ended tasks. These tasks comprise 15 research design assignments developed by small groups of three-four university students within the laboratory component of an academic course in Educational Research Methodology. The assignments were assessed by the teacher using an analytic rubric comprising six criteria and four performance levels with explicit descriptors; final results were graded on the Italian thirty-point scale. The underlying hypothesis is that ChatGPT, instructed to the same rubric, would produce assessments of the students' work that are not significantly different from those provided by a human assessor, both at the level of individual criteria and in terms of overall performance. The study's findings, also discussed in comparison with similar researches, may help to clarify potential of GenAI in supporting teachers' assessment-related tasks, particularly in the field of open-ended academic work.

Formative Assessment I

16:45 - 17:15

Navigating the Path to Effective Learning: Progressions in Assessment and Feedback Literacy

V. Scherman[1], F. van der Kleij[2], R. Shakra[2]

[1]IBO, Netherlands
[2]Australian Council for Educational Research, Australia

Feedback is pivotal in formative assessment, enabling students and teachers to respond to learning evidence and make informed decisions regarding learning. However, scaling formative assessment poses challenges due to stakeholders' varying understandings and skills.

This research addresses the need for practical, research-informed guidance to support curriculum-aligned assessment and feedback practices within the International Baccalaureate (IB). Employing an educational Design Research approach, the study developed prototype progressions for teacher and student assessment and feedback literacy. It builds on previous work that informed the development of a formative assessment model comprising five interrelated phases, illustrating the interaction between teacher and student practices, with feedback central to learning.

A rapid review of contemporary literature identified key models and skills associated with assessment and feedback literacy. Based on the review's findings, foundational, intermediate, and advanced levels were articulated and refined through expert consultation. These progressions encompass cognitive, behavioural, and socio-emotional dimensions of feedback literacy, reflecting the IB's philosophy of student agency.

Designed to foster a shared language and consistent practice, the progressions aim to enhance assessment alignment between intended, implemented, and attained curricula. The development process and the resulting assessment and feedback literacy progressions will be presented, highlighting key implications for their application.

Learning Oriented Assessment at an institutional level: Key design and validity prerequisites for positive impact

A. Salamoura[1]

[1]Cambridge University Press & Assessment, United Kingdom

In recent years, Learning Oriented Assessment (LOA) has emerged as an influential approach in the area of classroom-based assessment and beyond (e.g. Gebril, 2021; Leung et al., 2018; Purpura and Turner, in press). Although the advantages of LOA within the classroom have been extensively discussed, less attention has been devoted to its application in a wider context: at the school, university or national education system. Yet, recent models of learning programmes highlight the importance of viewing assessment as part of the broader learning programme they sit in (O'Sullivan, 2020) for understanding the conditions that may lead to their successful or unsuccessful application.

This paper presents a critical review of research on LOA conducted at institutional and national education levels (e.g., Gebril, 2021; Ho, 2015; Khan & Hassan, 2021; Leung, 2020; Qi, 2005), aiming to explore the factors that affect its validity. Four key prerequisites for ensuring LOA validity were identified: establishing educational coherence (cf. Wang et al., 2024), fostering an understanding of LOA principles among all key stakeholders (not just teachers), building collective teacher efficacy (Hattie, 2015), and closing skill gaps for teachers.

I will conclude by identifying areas in need of further research to complement the existing LOA literature.

School leaders' and teachers' attributions in making sense of norm-referenced school performance data

G. Molenberghs[1], R. Van Gasse[2], J. Vanhoof[2], S. De Maeyer[2], E. Goffin[3]

[1]Antwerp University, Belgium
[2]Universiteit Antwerpen, Belgium
[3]University of Antwerp, Belgium

School performance data can inform school improvement. However, data-informed decision-making is a complex and fundamentally interpretative process. This study examines school leaders' and teacher' attributions in making sense of their norm-referenced school's performance on the Flemish central reading comprehension test. Our focus within these attributions lies on the dimensions of locus of causality and controllability. Using a qualitative research design, we conducted 24 semi-structured interviews with school leaders and teachers. We applied framework analysis, expanding and refining the coding scheme through iterative open and axial coding. Attributions were categorized across four attributional levels: test characteristics, student characteristics, teaching practice and school practice. Findings reveal a dominance of external attributions, although internal factors were recognised as well, particularly at the teaching and school practice levels. Controllability emerged as a key factor for follow-up actions decisions. Furthermore, examination of attrbutional differences between population-referenced and group-referenced perspectives reveals that the latter tend to foster internal and external controllable attributions. By shedding light on the role of attribution processes, this study advances the broader discourse on data-informed decision-making for school improvement both theoretically and practically. Directions for further research are addressed.

e-Asssesment III

16:45 - 17:15

What examiners know: to what extent is detailed content knowledge necessary for reliable marking in broad, open-ended assessments?

M. Galache Ramos[1], M. Wami[2]

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

Summative assessments that comprise open-ended tasks, such as concept-based essay questions or coursework portfolios, are often designed to accommodate broad, flexible curricula that allow schools to select and study different content that is most appropriate for their local contexts. However, the flexibility that this breadth of content offers, in combination with examiners' varying degrees of familiarity or expertise with the specific content, is perceived to pose a risk to marking reliability as well as create manageability challenges to the standardization and marking process. In autumn 2024, an international awarding body embarked on an empirical study with an aim to explore the extent to which prior detailed examiner knowledge of curriculum content impacted marking reliability in broad, open-ended assessment tasks. This paper will present high level evidence of the results of a marking trial conducted in three assessment components specifically designed to investigate whether marking reliability depends on examiners' familiarity with the specific content or other variables related to the examiners' backgrounds. The paper will also present findings from the themes emerging from the analysis of the qualitative data collected during three focus groups which suggest methods to improve the guidance used to train and support examiners during marking.

Shifting Landscapes for Screening of Reading Difficulties in the Early Grades

B. Walgermo[1]

[1]University of Stavanger, Norway

Reading and writing difficulties are a major reason for referral to special education in Norwegian schools, often due to lack of early intervention. Although tools exist to help teachers identify struggling readers, their use is debated, and policy changes are made rapidly. Addressing this gap, this study first examines the teacher assessments compared to results from Norway's mandatory national reading assessment, and next, presents the development process of a new national adaptive screening test to meet teacher's needs.
In the first part of the study, 31 teachers identified students they believed needed support. They correctly identified 37% of students flagged by the screening test when including only those they were certain about, and 63% when also including those they were unsure of. Agreement between teacher judgment and test results varied widely—from 0% to 100%. The second part documents the design of an adaptive reading assessment used with nearly 50,000 Norwegian third-grade students. Key aspects include subtest selection, item formats, and adaptive model design, framed within a modern understanding of test validity.
The findings highlight the need for standardized assessment tools to support teachers' decisions and point to the potential of nonintrusive tools for monitoring both reading development and student motivation.

Typing Speed as a Proxy for Digital Literacy in Onscreen Assessments

S. Ishaque[1], E. Barrow[2], A. Ulicheva[3]

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

As onscreen assessments become increasingly common in educational and professional contexts, understanding digital literacy and its influence on performance grows in importance. Traditional evaluations often rely on self-reported questionnaires, which may lack precision and ecological validity. This study examines a task-based measure of typing speed as a proxy for digital literacy, aiming to understand its potential to modulate performance in onscreen assessments.

Building on Leijten & Waes (2013), our task accounts for critical language-specific factors such as lexicality, frequency, and keyboard layout. We explored its utility in two cohorts (n = 60 each) of adult L2 English speakers hypothesized to differ in digital literacy. Each participant also completed a high-stakes language proficiency test. Our research addresses three central questions: (1) Does the adapted typing speed task function as an indicator of digital literacy? (2) Does it predict test performance in an onscreen assessment context? (3) Does typing speed vary with test-taker ability?

Preliminary findings suggest that typing speed captures meaningful variance in test-taker behaviour. Our results affirm that most test-takers today operate with a baseline level of digital fluency compatible with onscreen formats. They also support the use of direct digital fluency measures when designing fair, accessible digital assessments.

National Tests and Examinations II

16:45 - 17:15

“It didn’t work well for me, but it works overall”: Exploring the relationship between validity and trust in a university entrance examination

P.J.(. Ho[1]

[1]University of Oxford, United Kingdom

This study investigates the relationship between validity and trust in the Hong Kong Diploma of Secondary Education Examination. Using qualitative data from a larger mixed-methods doctoral project, the study includes semi-structured interviews with 38 teachers and small-group discussions with 55 test takers. Initial findings suggest that the relationship between validity and trust is influenced by how stakeholders perceive the exam's primary purpose. For example, teachers raised concerns about the breadth-focused senior secondary curriculum, which they felt compromised the depth of learning. Although they recognised the challenge of assessing a wide range of skills while ensuring depth, their trust in the exam’s ability to reflect student preparedness was weakened. Students, while questioning the fairness of the school-based assessment process, accepted the exam’s role as an effective selection tool. This reflects a pragmatic form of trust, where the exam’s fairness in differentiating peers for university admissions is valued, despite some students feeling that the system was not beneficial for them personally. The analysis suggests a nuanced relationship between technical and social concerns related to the exam and its selection function, and that assessment literacy may play a role in stakeholders’ ability to critique the exam’s limitations while recognising its value.

Developing qualifications to support the Curriculum for Wales and to promote positive backwash

D. Jones[1]

[1]WJEC, United Kingdom

The Curriculum and Assessment (Wales) Act 2021 established the statutory framework for a curriculum for 3 to 16-year-olds in Wales. The Curriculum for Wales is a flexible framework where schools are encouraged to develop locally derived curricula to enable their learners to become ambitious, capable learners; enterprising, creative contributors; ethical, informed citizens; and healthy, confident individuals.

Concerns around the tensions between standardised high-stakes qualifications and the flexible and aspirational nature of the Curriculum for Wales posed challenges when developing Made-for-Wales General Certificates of Secondary Education (the main qualifications taken at 16). These included:

- specified standardised content being at odds with the concept of locally derived curricula

- potential for the qualifications to narrow or supplant the curriculum

- negative backwash effect of assessment on teaching and learning

- how to embed non-subject specific aspects of the Curriculum for Wales into qualification design in order to support schools in meeting their obligations in terms of delivering all the Curriculum.

WJEC produced principles and guidance to inform our qualification development work, ensuring that consideration was given to these tensions and trade-offs throughout our development process. This work will help others develop qualifications for positive backwash in 21st century qualifications.

Assessment of Practical Skills II

16:45 - 17:15

Foundations for Success: Reimagining Accessible Assessment Across the Years

M. Medhurst[1], J. McMillan[1], G. Rollo[1], K. Richardson[1]

[1]Australian Council for Educational Research, Australia

Assessment research and practice have traditionally adopted one-size-fits-all approaches, aiming for standardisation rather than individualisation. However, such approaches tend to treat students equally, not equitably, creating barriers to those who may not fit a uniform standard. Drawing on case studies from past and current research projects, this paper emphasises the critical importance of accessible assessment practices from early childhood education through to secondary school contexts. Findings highlight how validity of evidence across different layers of education should underpin student learning, school improvement and decision making across settings. Therefore, practitioners and education leaders need to examine accessibility for diverse groups of students across assessment design, implementation and interpretation of its results. In the early years context, the paper illustrates how learning progressions can help early childhood educators assess and support the development of diverse children's capabilities, reporting on practical design challenges in developing accessible assessment aligned with progression constructs Beyond preschool settings, we examine how an evidence-informed school improvement framework enables leaders to critically examine assessment data—and decisions based on these data—through an accessibility lens. The paper provides evidence-informed insights into developing accessible assessment across years of schooling, sharing valuable lessons learned for application across international contexts.

Measuring Literacy Skills for Young Learners: A School Grade-Based and Pluriliteracies-Informed Approach

L. Katkeviča[1]

[1]University of Latvia, Latvia

This study explores a school grade-based and pluriliteracies-informed approach to measuring literacy skills in Grades 1–3. Traditional early literacy assessments often focus on isolated reading skills, overlooking how young learners construct meaning across diverse formats and communication modes. Drawing on the Pluriliteracies for Deeper Learning model, the research introduces a trajectory-based assessment framework built around three core dimensions: skill type (e.g., information retrieval), subject-specific context, and cognitive depth using SOLO taxonomy. We developed grade-level learning trajectories and corresponding indicators, which guided the design of literacy tasks. These tasks were then assembled into diagnostic tests and piloted in three schools. The results highlighted three main challenges: interpreting visual texts, transferring information between written and visual modes, and combining information across different text types. Difficulties also appeared in keyword identification, summarizing, and drawing conclusions. The grade-based structure successfully supported class-level differentiation and laid the foundation for future development of individualized literacy trajectories. At this stage, the framework offers practical insights into measuring early literacy but requires improved task-indicator alignment and scoring reliability. Further research is needed to examine how learners progress along literacy trajectories over time.

Assessment Cultures III

16:45 - 17:15

Effectiveness of using open-ended items in high-stake examinations

N. Abdurahmanova[1], S. MIrzayeva[2]

[1]The State Examination Center of the Republic of Azerbaijan, Azerbaijan
[2]the State Examination Center of the Republic of Azerbaijan, Azerbaijan

This study explores the effectiveness of open-ended items in Azerbaijan’s school-leaving examinations from 2022 to 2024, introduced as part of a reformed national curriculum emphasizing higher-order thinking skills. The study aims to determine the extent to which open-ended items contribute to overall student performance and their ability to assess cognitive competencies beyond factual recall.
Using a quantitative approach and multiple linear regression, the analysis compared students' performance across multiple-choice and open-ended formats. Data from over 378,000 students were analyzed across three years. Results indicate that scores from open-ended items significantly and positively predict total exam scores, with R^2 values between 92.81% and 95.08%. Written responses were especially effective in identifying students with strong analytical and application skills.
The increasing predictive power of open-ended items over time suggests their growing relevance in fair and comprehensive assessment. These tasks support deeper measurement of students’ reasoning and communication abilities and should be further integrated into large-scale assessments.
Future efforts should focus on improving item design, exploring international practices, and leveraging technology for scalable, reliable scoring of open-ended responses to further enhance educational measurement and equity in high-stakes examinations.

The transition to the new SEC model. Did it succeed in providing a learner-centred model or did a one-size fits all model remain in place?

D. Pirotta[1], M. Micallef[2], F. Zammit[2], L. Vassallo[2], A. Grixti[2], J. Muscat[2]

[1]UNIVERSITY OF MALTA, Malta
[2]University of Malta, Malta

The 2025 assessment cycle marks Malta's first full implementation of the revised Secondary Education Certificate (SEC) model, aimed at shifting from a traditional, exam-heavy approach to a more learner-centred system. This reform aligned syllabi with the Learning Outcomes Framework and introduced assessment criteria designed to support clarity, structure, and differentiation. Instead of relying solely on end-of-cycle exams, the new model integrates school-based assessments over three years, culminating in a final exam with levels (1, 2, or 3) tailored to diverse learner abilities.
Using a mixed-methods approach, it analyses exam results, candidate feedback, and interviews with exam panel members. It examines alignment between learning intentions and outcomes, as well as students' understanding of assessment expectations. The results will inform recommendations for the refinement of subsequent examination sessions, with the goal of effectively bridging the gap between policy intentions and effective assessment. By identifying both strengths and areas needing improvement, the research will contribute to the ongoing evolution of assessment practices in Malta, aiming for a model that genuinely centres learners' needs, abilities, and growth trajectories.

Assessment of Transversal Competences in Vocational Education and Training. Theoretical and Practical Considerations.

D.L. Gray[1], D.H. Debowski[2]

[1]Independent researcher, United Kingdom
[2]Central Examination Board of Poland, Poland

Transversal competences, often called ‘21st century’, or non-cognitive skills, are viewed as vital for life and work. These skills, which include social and personal attributes, are essential across work and life activities and are increasingly incorporated in vocational educational and training programmes. Assessment of transversal competences influences both teaching priorities and student focus,and is crucial. Formative assessment helps in continuous feedback and development, but it may not ensure the recognition of these skills' importance. Summative assessment, linked to official credentials, might highlight their value but is challenging due to the complex nature of these competences. Transversal competences encompass a complex interplay of attitudes, values, behaviours, and skills that are difficult to quantify and standardise. This inherent complexity raises questions about the validity, reliability, and practicality of summative assessment methods in capturing the multifaceted nature of these competences. This presentation will discuss why and how transversal competences should be assessed, drawing on policies and practices from across Europe. It will report on a study that explores effective assessment methods and offers practical recommendations for policymakers to integrate robust assessment of these skills into vocational education and training programmes, ensuring they are viewed as integral rather than optional.

Poster Presentations

10:45 - 12:00

Aligning National Mathematics Curriculum and West African Senior School Certificate Examination Questions: Experiences from Ghana and Nigeria

A.P. Bah[1], D.A. Osoba[2], M.K. Dankwa[1]

[1]The West African Examinations Council Headquarters, Ghana
[2]The West African Examinations Council International Office, Nigeria

This paper examined the conformity of the West African Senior School Certificate Examination questions to the learning objectives. Subject experts were empaneled to review test items in terms of content, depth of knowledge and contextual characteristics. The study adopted content analysis design. Purposive sampling technique was used to select General Mathematics and Mathematics (Core), which is one of the requisite subjects for admission into tertiary institutions in both Ghana and Nigeria. For the sake of recency, 2022 and 2024 questions were use. Two rating scales, Forms A and B, and a Proforma were developed, validated and used to rate the questions. “Form A” was used to rate the extent to which the questions conform to the curriculum. “Form B” was used to rate clarity, simplicity and contextual characteristics of the items. The proforma was to classify the depth of knowledge of test items using revised Blooms Taxonomy. In each country, four-man subject panel was constituted to scrutinize the questions using relevant documents to make their judgement. Mathematics papers comprise two papers, multiple-choice and essay questions. The results from the two countries were compared, findings discussed, and recommendations made on how to sustain the best practices in test development procedure.

Participant ecologies: how the unique ecology of every individual affects their engagement with and responses to my research questions about the impact of qualifying as a Chartered Educational Assessor

M. Talbot[1, 2]

[1]University of Leeds, United Kingdom
[2]Chartered Institute of Educational Assessors, University of Hertfordshire, United Kingdom

This poster will report on one of five case studies of teachers in English secondary schools that form part of my doctoral thesis. I am aiming to better understand my participants' ecologies, to build rich pictures, and elucidate how their unique circumstances have led them to where they are now in their educational assessment journeys. Building on the work of Bronfenbrenner (1974, 1977) and Bronfenbrenner and Ceci (1994), I will present some interim findings, including my model of participant contexts.

Training to become a Chartered Educational Assessor forces successful CEAs to reflect on what it means to be an educational assessor, to review their self-understanding, self-efficacy, identity, and impact (Lortie, 1975). They re-negotiate and re-interpret their beliefs and perceptions according to their individual motivations and ecological context, mediating their responses and actions accordingly (Podesta and Hoath, 2020). CEAs demonstrate a moral commitment to leading and influencing educational assessment well beyond the technically competent, integrating it with learning, teaching, and pedagogy, to encompass a broad range of cognitive and emotional dimensions in themselves and their colleagues (Kelchtermans et al., 2011). These complex and multi-faceted emotional and psychological internal dialogues of the participants are important dimensions of their overall ecology.

Developing students' argumentation skills through situational tasks in preparation for external assessment

L. Kabakbayeva[1], B. Abylkhatov[1, 1]

[1]NIS PhM school in Aktobe, Kazakhstan

Developing students' argumentation skills is an important goal of modern education. The ability to formulate and justify one's point of view, analyze mistakes and draw conclusions plays a key role in academic success and preparation for external assessments. The PISA International program has revealed a problem: Kazakhstani students are proficient in subject knowledge, but have difficulty applying it in real situations. One way to solve this problem is through situational tasks that help students analyze data, hypothesize, and build arguments.
The study assessed the impact of situational tasks on the development of students' argumentation. Experimental and control groups were formed, in which the results of the tasks were compared. The methodology included the development of educational materials, conducting lessons with situational tasks, analyzing student responses and collecting feedback.
The results showed that the use of situational tasks contributes to the development of logical thinking, improved argumentation and increased student engagement. The teachers noted their active participation in the discussions and motivation to learn. Thus, situational tasks are an effective tool for integrating theoretical knowledge with its practical application, which helps prepare students for academic and professional challenges.

Facilitating Learner Reasoning: Digital Annotation Tools for High-Stakes Assessment

A. Khan[1]

[1]Cambridge University Press & Assessment, United Kingdom

Research by Cambridge University Press & Assessment (CUP&A) demonstrates that learners actively engage with examination items rather than passively reading them. On paper assessments, learners highlight keywords, strike out improbable multiple-choice options, and mark up diagrams to create connections between different components. Our in-depth classroom observation and follow-up interviews with learners indicate these annotation practices aid their comprehension and recall, suggesting that the absence of such tools in digital exam platforms could negatively affect candidate experience and potentially compromise test validity.

This poster will outline the development of an annotations toolkit designed to aid learners' cognitive processes during digital examinations. The toolkit—developed through collaboration between assessment researchers, practitioners, UX designers, and developers—follows an iterative product development approach involving rapid prototyping and continuous testing with learners. The toolkit currently includes functionalities such as a rule-out tool, highlighter, and image markup tool, with future enhancements such as an equation writer also planned.

Ultimately, the aim of this toolkit is not merely to replicate paper-based annotation processes, but to deliver a digital solution that minimises construct-irrelevant variance and enhances overall candidate performance by providing a streamlined exam experience in which learners can focus purely on demonstrating their content knowledge.

Students with disability as research advisors: Exploring fairness and justice in university assessment

K. Dyliaeva[1]

[1]Queen's University Belfast, United Kingdom

This poster will detail the 1st phase of my PhD project, which has been guided by student advisors in designing research strategy and questions. In this presentation, I will demonstrate how student voice and lived experience of disability may enrich the investigation of assessment processes in a university in the UK. Few assessment studies recognise students' competence as research advisors and rarely adopt their thinking about assessment. This study uses a Student Advisory Group (SAG) of undergraduate students and implements their advice in research practice. This matters because students with disability are directly impacted by the processes underlying assessment practice and thus have a more comprehensive experiential view of disparities in assessment events. The SAG argued that assessment for them has been nurtured by neurotypical and predominantly able-bodied expectations of students. This guided me to take a sociocultural lens toward assessment and extend focus from students' views of fairness to lecturers' assumptions about students' abilities. Thus, the study aims to advance our understanding of fair assessment practice in higher education in the UK and answer the question: What does fair and equal assessment look like from the experience of undergraduate students with disability in a mainstream university in the UK?

Designing Inclusive Assessments: Strategies for Equity and Accessibility in Education

D. Tsagari[1]

[1]Oslo Metropolitan University, Norway

Designing inclusive assessments is critical to creating equitable and accessible learning environments that serve the diverse needs of all students. This poster presents a framework grounded in Universal Design for Learning (CAST, 2018) and differentiated instruction (Tomlinson, 2001), offering practical strategies for reducing barriers to assessment for neurodiverse learners, students with disabilities, and those from culturally and linguistically diverse or marginalized backgrounds. It highlights the transformative role of digital tools in enabling personalized assessment pathways and enhancing accessibility through assistive technologies and flexible formats. Formative feedback is emphasized as a mechanism of empowerment, helping students monitor progress and engage in self-directed learning (Hattie & Timperley, 2007). Real-world case studies illustrate the implementation of inclusive practices across educational contexts, showing how thoughtful assessment design fosters fairness and deeper engagement.
Key challenges, including systemic bias, lack of institutional responsiveness, and professional development gaps, are critically examined with recommendations for addressing them through sustainable policy and practice. This poster promotes inclusive assessment as central to designing tomorrow's assessment landscape for positive impact on learning ensuring that assessment not only measures achievement but also enhances opportunity, motivation, and equitable educational outcomes.

Washback effects on teaching practice: the Greek Language University Entrance Examination

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

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

This poster investigates how high-stakes examinations shape classroom practice, focusing on the washback effects of the Greek Language University Entrance Examination. Drawing on existing washback and validity frameworks, the study examines how test format, content, and implicit expectations influence pedagogy at the micro level. Content analysis of teacher reflections and lesson recordings (a total of 20 semi-structured interviews and transcripts of 10 lesson hours) reveals a narrow instructional focus on writing skills such as grammatical accuracy, memorised vocabulary, and rigid stylistic forms—at the expense of genre awareness, argument structure, and critical thinking. Despite intensive preparation, students demonstrate minimal improvement in these key areas of literacy, reflecting a disconnect between assessment practices and meaningful learning. Echoing Messick's (1996) concern with construct underrepresentation, the university entrance exam fails to capture broader competencies such as critical literacy and authentic language use.

The study highlights the need to redesign assessment systems that integrate learning aims and reflect real-world language use. It also advocates for a shift away from decontextualised, test-driven pedagogy toward practices that promote deeper engagement, equity, and critical reflection. In doing so, it aligns with the broader goal of designing tomorrow's assessment landscape for positive impact on learning.

Grading Effort in Physical Education. A Norwegian Case Study

R. Korshavn[1], E.G. Gjølme[1], L.V. Sandvik[1]

[1]NTNU, Norway

Norway is one of few countries where physical education (PE) teachers are required to assess student “effort” as part of final grading—a policy that raises both pedagogical and ethical dilemmas around fairness and reliability. While intended to promote inclusive and holistic assessment, the concept of effort remains subjective and broadly defined in curriculum documents (Brookhart et al., 2016; Aasland & Engelsrud, 2017).

This qualitative case study explores how PE teachers interpret and apply this policy in practice. Drawing on focus group interviews with teachers from seven schools, and analyses of national and local assessment guidelines, we examine how effort is understood and operationalized in relation to other assessment criteria.

Findings reveal wide variation in teachers’ interpretations of effort—ranging from visible exertion to improvement, motivation, and contribution to group dynamics. Teachers express uncertainty about how to balance effort with skill development, reflection, and knowledge. These inconsistencies are linked to vague curriculum constructs, uneven assessment literacy, and unequal access to professional development (Jærnes, 2023).

We argue that meaningful and fair assessment in PE requires clearer curricular guidance, structured professional dialogue, and shared interpretive frameworks. The study contributes to international debates on the challenges of assessing non-cognitive factors in school subjects.

Cambridge Early Adopter Programme: building a pathway to digital exams

S. Kemmery[1]

[1]Cambridge University Press & Assessment, United Kingdom

This poster highlights Cambridge's Early Adopter Programme (EAP) as a strategic initiative trialling digital Exams across international centres ahead of full implementation in 2026. Drawing on the feedback from 31 EAP schools across Europe, MENA, SSA, Pakistan and the US in 2025, this study interrogates the pedagogical, infrastructural, and institutional dynamics shaping the digital exams landscape.
Findings demonstrate strong stakeholder engagement and positive learner experiences are key to digital adoption; however, key constraints—such as BYOD limitations, the need for hybrid exam administration, and concerns around technical resilience—underscore the complexity of system-wide readiness. In response, Cambridge is pursuing a phased and scalable strategy grounded in customer feedback: the expansion of familiarisation tools, the communication of updated exam administration guidance and a comprehensive training and support model for educators and invigilators.
By integrating short feedback loops into the design cycle, this work illustrates how assessment reform can serve teachers and students, ensuring that these audiences are at the heart of development that can provide examples such as improved accessibility features and accessing resources on the platform such as the periodic table. The project emphasises inclusivity, and stakeholder trust, emphasising the importance of co-design in shaping the future of assessment.

Predictive Validity of Digital Mock Assessments

C. Vidal Rodeiro[1], T. Gill[1]

[1]Cambridge University Press & Assessment, United Kingdom

Mock exams are low-stakes assessments designed to mimic the format, content, and conditions of the examinations students are preparing for and can have a big impact on how teaching and learning are approached in the classrooms. In particular, they can help teachers know the strengths and weaknesses of their students and identify areas of content which need greater emphasis or clarity. Mock exams can also inform teachers of how students might do in live examinations.
In the context of transitioning from paper-based to digital assessments, Cambridge University Press & Assessment offers a Digital Mocks Service for schools in England and around the world. Using data from assessments delivered via this service (in qualifications such as Computer Science, English Language, Psychology), we investigated how well performance in digital mocks predicts performance in live exams. The digital mocks were based on exams delivered on paper in previous live sessions or on practice/sample papers. The research findings show high correlations between mock and live assessments, indicating good levels of predictive validity in most qualifications. The predictive validity of mock exams is important as it helps to validate their use as a diagnostic tool that can have a positive impact on learning.

Assessing consistency in exam standards across sessions

M. Wami[1], A. Furlong[1]

[1]International Baccalaureate, Netherlands

The International Baccalaureate (IB) conducts two exam sessions per year for its Diploma Programme (DP), in May and November. Exam standards must remain consistent across sessions to ensure fairness and credibility. Variations in difficulty between sessions can unfairly affect candidates, impacting their academic progress such as university admissions. This poster will present ongoing research, investigating whether the overall standards of the IB's May and November DP sessions are comparable. Using historical exam data, analyses of the change in performance of students retaking exams in the session immediately following their first attempt will be conducted to assess whether one session is consistently easier or harder. First, cluster analysis will be performed to classify retake students into subgroups, representing their different characteristics. Secondly, regression modelling will be conducted to evaluate whether there were changes in performance between sessions, controlling for other potential factors that may influence grade variations. The resultant cluster membership from the first step will also be included as a predictor in the model. This study will provide valuable insights into exam consistency by analysing retake performance. The results will inform IB policymakers on whether adjustments are necessary to maintain equity in assessment across sessions.

Utility of expert judgement for setting grade boundaries in England

A. Miranda[1], T. Benton[2]

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

High-stakes assessments impact students' paths; therefore, setting high-quality standards is essential. Each year, grade boundaries are set using statistical evidence and subject expert judgement. Experts review marked scripts to decide whether scripts accurately represent the corresponding grade following a 'top-down bottom-up' approach.

The study statistically examines these judgements for grade awarding, as there are no recent reviews of their value. The analysis reviewed 676 grade boundary decisions from 2024. Most used three experts who each reviewed two scripts per mark.

We analysed various ways expert judgement could be used in awarding. First, we explored different approaches to defining a zone of uncertainty within which the real boundary should lie. We found that certain (highly plausible) approaches failed to generate a zone at all almost half the time. However, under the best definition, the real grade boundary was within the zone 86% of the time. We also explored formal statistical approaches to generating boundaries from expert judgements. These were within 1 mark of operational boundaries 71% of the time. However, they were often more generous than operational boundaries for grades relating to high levels of performance. The study concludes with a discussion of the utility of expert judgement for awarding.

Sociodemographic characteristics, prior attainment and university choices in the UK: an intersectional approach

D. Tonin[1]

[1]Ofqual, United Kingdom

Students' grades in their school-leaving qualifications are the most important determinants of progression to university in the UK. Differences in attainment between groups of students can have multiple and complex causes and have implications for university participation. Earlier research looked at the effects of sociodemographic characteristics and prior attainment on the decision to apply to university independently, not considering the heterogeneity across the concurrent experience of different combinations of individual characteristics. In recent years, there has been a call for applications of intersectional approaches in educational research to address the multidimensional impact of students' characteristics on educational inequalities. Drawing on intersectional theory and using a novel multilinear regression technique called intersectional multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA), I explore how the students' decision of applying to university varies along the intersecting axes of different students' characteristics. Results indicated that students' characteristics significantly impact students' decision to apply to university, with additive effects being more prominent than interactive ones. The presence of intersectional effects suggests it is important for those within the school and the university system to consider how some students may be multiply marginalised and face heightened challenges in progressing successfully through their educational journey.

The effectiveness of group discussion, assessment criteria, peer assessment in improving the geography essay writing skills of 11 grade students

M. Yeskeldi[1], M. Kurbangaliyev[1]

[1]Nazarbayev Intellectual school Aktau, Kazakhstan

This research explores the effectiveness of group discussion, assessment criteria, peer assessment in improving the geography essay writing skills of 11 grade students. This study lasted 10 weeks. Pretest, posttest, interviews were employed among control and experimental groups. Writing essays in the External examination of geography is a challenge for high schoolers. To write a well-developed cause and solution essay, not only should they be able to structure the essay correctly, use geography terms properly but also they need to think critically to find solutions and explain the reason for their choices and consequences of their resolution. Originality of solutions is a key criterion in the assessment. Therefore, students discuss their strategies and listen to their peers' ideas , give and receive feedback according to the assessment criteria. They need to write an essay that has meaningful content , is grammatically correct, and follows the spelling rules and punctuation. Peer assessment allows students to notice mistakes and improve them. It enhances students' critical thinking, and confidence. Results illustrate that discussing ideas, peer assessment, following assessment criteria leads to using geography terminology accurately and immense improvement in the structure and the quality of the solutions and examples.

Trends in between-school and between-class variation in student achievement in the Nordic countries based on IEA assessments from 2011 to 2023.

S. Määttälä[1], S. Kupiainen[1], J. Christensen[2], A. Christensen[2]

[1]University of Helsinki, Finland
[2]Aarhus university, Denmark

This study explores the development of between-school and between-class variation in student achievement in Nordic basic education between 2011-2023, drawing on international large-scale assessments conducted by the IEA: TIMSS and PIRLS. Recent demographic and geographical developments, for example in Finland, such as increasing residential segregation, have contributed to growing disparities in student backgrounds across schools, reflected in learning outcomes from an early stage.
We address the following research questions: (1) How has between-school and between-class variation evolved in Nordic countries excluding Iceland? (2) To what extent can differences in achievement be explained by structural factors at the school and class level compared to individual socioeconomic status? (3) How have these associations changed between 2011 and 2023?
Our analytical framework is based on the Douglas Fir Group's (2016) transdisciplinary model of language learning, which considers how learning is shaped by individual, social, and structural factors. Using multilevel modeling and decomposition analysis, we examine how much of the variance in student achievement is attributable to school-, class-, and individual-level characteristics. Control variables include socioeconomic status (SES), absenteeism, and language spoken at home. Findings show increasing variation in Finland and Sweden, particularly in reading, while Norway and Denmark show more stable trends.

In the eye of the digital storm: A blind spot for fairness in screen-based learning?

C. Vincent[1]

[1]AQA, United Kingdom

While there are benefits associated with on-screen assessment, it is important to consider its potential effects on student health and inclusivity. Moving to digital exams will result in more students spending prolonged periods in front of screens. This poster brings into focus how increased screen exposure (particularly 415–455 nm blue light) contributes to digital eye strain, disrupted sleep, myopia and age-related macular degeneration.

Research shows that symptoms including eye fatigue, blurred vision and headaches are common among students and exacerbated by close-range screen use and poor ergonomics. However, appropriate screen sizes and improved positioning have been found to mitigate these symptoms during prolonged screen exposure. Interventions like the 20-20-20 rule (20-minute break to view something 20 feet away, every 20 minutes) can potentially reduce eye strain but may not be practical in high-stakes exams. Furthermore, commercial solutions like blue-light filters require independent evaluation.

This poster highlights the importance of research-led guidance around screen use and eye health. Inclusive, evidence-informed policies that limit blue-light exposure, support ergonomics and encourage diverse assessment design could help protect students' visual and general wellbeing. As the education sector continues to embrace digital delivery, turning a blind eye to ocular consequences is not an option.

From pen to keyboard: Is fairness affected by mode of input?

F. Walker[1], S.A. Husain[1], C. Vincent[1], V. Armstrong[1]

[1]AQA, United Kingdom

With digital exams on the rise, it is important to consider whether all learners are treated fairly. The mode of input—typing versus handwriting—may affect students' performance, depending on their typing fluency and handwriting legibility. This poster presents findings from a literature review exploring how mode of input affects diverse learners, particularly those with special educational needs and disabilities.

For some students, especially those with handwriting difficulties, typing improves legibility and reduces cognitive load, enabling fairer outcomes. Typed responses also allow easier editing and proofreading, which may enhance performance. However, students with limited typing fluency or motor skills—such as some students with dyslexia—may struggle to fully demonstrate their knowledge.

The mode of input also has implications for marking, raising concerns about assessment validity when work is graded not on understanding or knowledge, but on presentation. While typed responses are easier to read, they may be marked more harshly due to visible errors and higher expectations for presentation quality.

Evaluating students' typing fluency and building digital skills are key to making digital exams fair and inclusive for all learners. This poster also invites discussion on how exam tasks and marking can support equitable outcomes.

Simulating Expert Behavior with LLMs for Question Difficulty Estimation: Angoff-like procedure versus Pairwise Comparison

D. Kolesnikova[1], K. Fedyanin[2], M.J.S. . Brinkhuis [3], M. Bolsinova[1]

[1]Tilburg University, Netherlands
[2]Yandex, Serbia
[3]Utrecht University, Netherlands

As large language models (LLMs) become increasingly integrated into education and assessment, their potential to support test development processes remains underexplored. This poster presents an early-stage study investigating whether various LLMs can simulate expert judgements of item difficulty for the multiple-choice tasks.
Two procedures were explored: (1) a simplified Angoff-style estimation of the probability that a typical student would answer correctly, (2) pairwise comparisons of item difficulty. LLM-generated estimates were compared to empirical student performance data. Experiments showed a moderate positive correlation with real student data suggesting that LLMs can meaningfully capture relative item difficulty under certain conditions.
Other conditions that were studied are: prompt design, item order, few-shot learning. Study introduces a few novel datasets as well as ones used before. The findings highlight both the potential and limitations of using LLMs for psychometric purposes, including scalability and sensitivity to procedural framing.
Future work will explore how LLMs handle open-ended items and whether combining elicitation procedures can improve accuracy. This study contributes to the emerging conversation on AI-supported assessment design.

Stakeholders’ Perceptions on Assessment Systems in Secondary Education.

J. Godschalk[1, 2]

[1]Maastricht University, Netherlands
[2]Cito b.v., Netherlands

Assessment is a critical part of education, where evidence of learning is systematically collected to draw inferences about students’ progress (Harlen, 2004). Assessment results in informed judgement and decisions about students. Some of these decisions are summative, such as determining educational levels, grading students’ performance and awarding qualifications (Howard et al., 2021). Conversely, others are formative decisions that nurture learning. Both decisions can have multiple consequences and potentially influence students’ future education and careers. Therefore, the quality of assessments is paramount; they must be of indisputable high quality to ensure satisfactory judgement and decision-making.
Quality of assessment in general is defined by the quality of instruments, procedures, policy, organisation and skills of stakeholders involved in assessment (Joosten-ten Brinke, 2011). In higher education, knowledge exists pertaining to the realization and improvement of assessment quality. However, such understanding remains lacking within the context of secondary education. This quantitative study aims to identify stakeholders' perceptions of assessment systems’ quality in Dutch secondary education schools, by using a questionnaire. The relative importance of quality criteria will help schools identify their strengths and weaknesses. With this information, it is hoped to help schools ensure high-quality assessment and sound decision-making.

Efficiently and effectively increasing the role of formative assessment in large-scale programs

A. Kramer[1], Y.P. Hsiao[1], M. Stienstra[1]

[1]Tilburg University, Netherlands

Formative assessment and feedback are essential for student learning, but their impact is significant only when thoughtfully designed. In large-scale educational settings, several contextual factors complicate formative assessment design including large-class sizes, workload for teachers, students' attitudes toward formative assessment. This study aims to: (1) synthesize findings from the literature to identify program-level design principles that can enhance the role of formative assessment and feedback in large-scale business education, and (2) analyze current formative practices and evaluate the extent to which they align with these principles, ultimately proposing an improvement plan for the annual quality assurance cycle. We adopted a mixed-methods approach. We conducted a scoping review to identify contextual barriers, design principles, and effective strategies for implementing formative assessment and feedback in large-scale business education settings. We analyzed data on formative assessment and feedback practices across 21 programs at a large business school in the Netherlands. Initial results show that there is little research on formative assessment in large-scale programs and that most research is on course-level and not of the program-level. In addition, the formative assessment loop is not always closed, meaning that feedback is given, but students are not encouraged to use the feedback.

Lost in transcription: Why accuracy matters in AI marking

K. Gilmartin[1], L. Howarth[1]

[1]AQA, United Kingdom

With the world of education and AI rapidly evolving, we have been exploring how using AI in marking could improve marking efficiency and mitigate teacher workload. In England, pupil classwork is still predominantly handwritten; therefore, large language model input data is a conversion from handwritten text into digital format. But how much does transcription accuracy affect AI marking performance?

We investigated this by comparing the accuracy of various AI transcription tools against a manually transcribed sample of 100 student responses to exam questions. Using text similarity metrics, such as Jaccard, Levenshtein, Ngram and embedding-based comparisons, we calculated average transcription accuracy across the tools. The results found that the accuracy scores ranged from 78.2% to 92.2%.

To assess the impact of transcription quality on AI marking performance, we selected a GCSE English Language question and ran 93 student responses through an AI marking model. This process was conducted twice, using both the most and least accurate transcriptions. The results demonstrated that the AI's marking accuracy was notably improved when the more accurate transcriptions were used. These findings highlight the important role transcription accuracy plays in the overall effectiveness of AI assessment.

Authentic Learning Journeys through Collaboration

M. Stienstra[1], A. Kramer[1], Y.P. Hsiao[1]

[1]Tilburg University, Netherlands

Collaborative learning and group assessments have been extensively used in higher education over the past two decades. While these methods are often designed to simulate real-world environments, the concept of authenticity within program curricula and courses remains unclear and sometimes appears to be driven more by a desire to reduce teachers' grading time. Moreover, it is unclear how students navigate collaborative learning with peers throughout their courses and across the entire program. This study aims to analyze students' current learning journeys within a program that integrates collaborative learning and group assessments in courses. Specifically, we examined program learning outcomes related to collaborative competencies and the design of instructional and assessment strategies. Additionally, we interviewed students and teachers about their experiences. Based on these findings, we visualized the student learning journey and evaluated its authenticity by exploring the relevance of tasks and their impact on students' self-identity, self-esteem, and overall well-being. We propose improvements and a program-level framework for authentic collaborative assessment design, focusing on social value and learning outcomes. This framework aims to strengthen the development of students' knowledge, skills, work readiness, and personal growth by promoting collaborative learning throughout the entire program.

The Impact of Accessibility Features in GCSE Science Assessments

E. De Groot[1]

[1]Cambridge University Press & Assessment, United Kingdom

This study looks at how accessibility principles have been built into GCSE Science assessments in the UK education system for students aged 14-16 years old and what impact they have had. The focus is on updates to language, visuals, structure, and content support in papers from 2018 to 2024, across foundation and higher tiers.

Findings show consistent use of simpler language, more effectively broken-down questions, and improved contextual support. Since 2022, there's also been a noticeable increase in the use of diagrams, improving visual accessibility.

Statistical analysis of item performance in these question papers shows that overall difficulty has stayed consistent. Specifically, this analysis looked at item facility, discrimination and how often students don't attempt an item. However, questions worth more marks are better at distinguishing between students of different ability levels—suggesting that accessibility changes haven't affected the difficulty of the exams.

These changes align with key accessibility principles such as consistency and familiarity, helping to remove unnecessary cognitive hurdles while keeping academic standards high. This research highlights how thoughtful exam design can make assessments more inclusive—offering valuable lessons for future practice and contributing to the broader goals of equity and fairness in education.

Antecedents of AI's Perceived Usefulness in Grading: Insights from Norwegian High School Teachers

H. Eriksen[1], E. Elstad[2]

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

Artificial intelligence (AI) has sparked challenges in educational assessment globally since the introduction of AI in late 2022, particularly due to its ability to create human-like text. A key concern is the validity and trustworthiness of test results, which are essential to the assessment system (Black & Wiliam, 2018). While there is optimism about AI's potential to reduce teacher workload through automated grading, there are also notable concerns (Mai et al., 2023).
This study aims to explore factors influencing teachers' perceived usefulness of AI for grading assignments and tests. A survey was conducted among 223 high school teachers in Norway, and we analyzed relationships between latent variables using structural equation modeling. We focused on how instructional self-efficacy, perceived time pressure, and the value of experience in teaching relate to the perceived usefulness of AI in grading.
Our findings reveal that teachers exhibit a cautious attitude toward AI in grading. There is a strong positive correlation between instructional self-efficacy and the perceived usefulness of AI, a weaker correlation with the importance of experience, and no relationship between perceived time pressure and AI's grading utility. This last insight was surprising, as reducing teacher workload is typically highlighted as a key improvement area.

Using longitudinal datasets to explore labour market outcomes for learners taking Vocational and Technical Qualifications in England.

B. Ashworth[1], H. Dalton[1], R. Bhatta[1]

[1]Pearson, United Kingdom

This poster reports on research looking at employment earnings for learners completing upper secondary vocational and technical qualifications (VTQs) in England. Using the Longitudinal Education Outcomes datasets (LEO), we can see where VTQs have the most impact in respect of median salary for learners going directly to work and for those accessing university before entering the labour market.

In the past decade, several studies, including The Nuffield Foundation (2022), have shown the impact of VTQs on supporting progression to Higher Education. Much less is reported about longer-term employment outcomes for learners that take VTQs in respect of salary outcomes and impact on employment pipelines in skill-shortage areas such as health and digital sectors.

Regression and multi-level modelling, along with descriptive statistics have been used to look at salary outcomes through interactions that are known to impact salary, these include: economic disadvantage, lower secondary attainment, gender, SEND status, and sector/subject studied. We also spotlight the employment sectors where learners taking these qualifications and progressing to university have added the most and least salary benefits to learners, intersecting these with learner characteristics where the matched data allows.

A Theoretical Framework for Digital-First Assessment: Validity, Reliability, and Accessibility

L. Liu[1], I. Custodio[1]

[1]Pearson, United Kingdom

As awarding organisations in the UK look to transform from paper-based and dual-mode delivery to digital-first assessments, questions are posed around how we evaluate their validity. Inspired by Bandalos's (2018) book, this poster describes a framework that seeks to integrate best practices from international large-scale assessments into the national context, emphasising reliability, validity, and accessibility as essential elements during the digital transition.

This poster introduces a structured six-step methodological tool designed to collect comprehensive research evidence for developing dual modes and digital-first assessments that are psychometrically sound and accessible, including: (1) Assessment Design and Construct Validity; (2) Content Validity and Consequences of Testing; (3) Usability and Accessibility; (4) Reliability Analysis; (5) Validity Argument Framework; and (6) Iterative Improvement. The poster clearly defines and details the evidence required for the initial four steps, subsequently illustrating how step five synthesises these findings into a cohesive validity argument. Finally, the sixth step emphasises iterations, informed by research findings and user feedback, to ensure ongoing enhancement of assessment quality.

Overall, the methodological tool provides a systematic and rigorous approach to designing and developing digital-first assessments that uphold high psychometric standards and accessibility.

The reliability of examiners' judgements of the likelihood of scripts being produced by background speakers of Japanese

P. Hallsworth[1]

[1]International Baccalaureate, United Kingdom

IB Diploma Programme students must take a course in Language Acquisition (e.g., in Japanese). These courses are for language learners, not for students who already speak the language. When marking students' oral exams, examiners flag students they believe to be 'misplaced' background (native) speakers. Examiners are not currently required to raise flags when marking written exams.
This mixed-methods study investigated the reliability of examiners' judgements of the likelihood that written work had come from background speakers. In a purposive sample of seven examiners of IB Japanese language courses, participants were asked to make judgements as to the likelihood of written scripts belonging to 'true' learners or misplaced background speakers of Japanese. Participants were also interviewed to explore the features of the scripts that led to their judgements and to discuss the issues of fairness and standards that student misplacement raises.
Participants' judgements were relatively reliable (ICC=0.89). Their judgements were partially validated by teachers' classification of candidates as first-language Japanese (or otherwise) at point of exam entry. The interviews revealed tensions between senses of (un)fairness, as well as a nuanced interpretation of what it means to be a language learner. A more holistic approach to identification of background speakers is recommended.

“I’ll get a dog if I get an A” – Rewards as a Parental Strategy to Support Students’ Academic Success

D. Normann[1], H. Fjørtoft[2], L.V. Sandvik[1]

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

Parents sometimes give material rewards to their children for earning high grades in school. While such practices may enhance external motivation (Deci & Ryan, 1985), they may also increase performance pressure (Chamberlin et al., 2018), potentially affecting students negatively. This study aims to explore the dynamics and consequences of rewarding students for earning high grades.

Drawing on open-ended survey responses from Norwegian upper secondary students (n=1,708) and 28 focus group interviews of 22 parents and 99 students, we explore students’ and parents’ perceptions of rewarding good grades. We emphasize parents’ rationales for rewarding high grades, the types of rewards given, and how students perceive these rewards.

Our results suggest that parents rely on material rewards due to a lack of alternative methods to motivate students. Rewards include money, travel opportunities, leisure activities, or other privileges. However, while rewards motivate students to perform, they also increase performance pressure. Thus, while rewards promote external motivation, they may also potentially undermine students’ well-being and learning. We discuss these findings in the context of existing research on intrinsic and extrinsic academic motivation in education and the broader impact of grading practices on student wellbeing and student-parent relations.

Peer Feedback? Maybe. Instructor Feedback? Definitely: Exploring Student Perceptions of Feedback in Higher Education

N. Hasan[1], F. Shamsaldeen[1]

[1]Kuwait University, Kuwait

Feedback is essential to student learning, as it helps them understand their current standing and how they can improve. This study examined how university students perceive two types of feedback: one from instructors and the other from their peers. We focused on four aspects: how trustworthy, clear, motivating, and usable the feedback felt.

Sixty-one student groups received two sets of feedback on their projects—one written by their instructor and the other by another student group. Without knowing which was which, they rated both using a 16-item survey. The results showed that students consistently saw instructor feedback as more trustworthy and clearer. While peer feedback was viewed as slightly less effective in motivating and helping with revisions, it still offered value, particularly in being relatable.

Students who correctly identified instructor feedback often pointed to its precise language, alignment with course standards, and a more professional tone. Peer feedback, while sometimes less polished, was appreciated for its conversational style and accessibility.

This study shows that while students may naturally lean toward valuing instructor feedback, there is room for peer feedback, especially when supported by clear training and expectations. Combining both could strengthen how feedback supports learning in higher education.

Test Equating for Maintaining Standards Using Comparative Judgement

K. Mason[1]

[1]Pearson, United Kingdom

In English schools, there are often several options available to students aiming to achieve the same qualification. Standards are maintained using statistical indicators. This poster describes an approach where additional evidence is gathered using comparative judgement to ensure grades awarded across two versions of a GCSE English qualification represent the same standard.
We consider two distinct versions of GCSE English, both meeting the same content and objectives. “English 2.0” engages students with contemporary texts, attracting older students and those re-sitting. Supplementary evidence is needed to support grade setting decisions. Equating the assessments of the two qualifications is essential, but traditional approaches are not applicable as there are no common questions.
In 2024, comparative judgement exercises were conducted using assessments from both qualifications. Judges compared pairs of assessments, with each piece of work part of 22 or 23 pairs. Judges with infit values greater than 1.2 were removed from the analysis. Separation reliability was around 0.96. Theta values were regressed against qualification marks. These theta estimations were bootstrapped in order to obtain a range of possible marks, in the regressions. The range found was around seven marks in the middle of the range.

Chatbots in Language Education: Challenges and Possibilities

Ø. Gilje[1]

[1]University of Oslo, Norway

The integration of artificial intelligence (AI) into language education has a long history, featuring various domain-specific pre-programmed chatbots (Şahin Kızıl et al., 2025). Recent advancements in large language models and generative AI now enable teachers as end-users (Author et al., 2025) to customize chatbots for specific roles, facilitating feedback that enhances students' abilities in both oral and written skills in second (L2) and third languages (L3).

This poster presents a case study from the innovation project "X" (2024-2027, funded by the Norwegian Research Council), which investigates how Generative AI and Google Workspace for Education support feedback and scaffolding in language education. Focused on a lower secondary school setting, the study explores how students engage with the chatbot Lingu during Spanish lessons. Drawing on extensive fieldwork over 15 months and interviews with students aged 15-16, the poster highlights how teachers can provide tailored prompts through customized chatbots to support student preparation for oral and written examinations.

By incorporating generative AI into assessment task design, educators can better facilitate authentic assessment practices that engage students critically with AI-generated content. The findings offer insights into designing formative assessment tasks for AI use in language education, addressing both challenges and opportunities.

Dealing with disruption: the evolving nature of teacher assessment practices in technology-driven formative assessment.

K. Menary[1]

[1]City and Guilds, United Kingdom

While some recent research has begun to explore the challenges faced by teachers in dealing with the rapid switch to online delivery due to the Covid-19 pandemic, there has been little research specifically focusing on how teachers adjusted their formative assessment practices over time, and how different constraints (and affordances) introduced by digital technology interacted with teachers' emerging digital assessment literacy.

To address this issue, I conducted a longitudinal, multiple case study of eight teachers working within Europe. In this presentation I will discuss three key themes: (1) issues with rapport in digitally-mediated environments that affect a teacher's ability to accurately assess learners; (2) the widespread use of technology not specifically designed for teaching and assessing; (3) the generation of reams of data which cannot easily be utilised to improve learning outcomes. The analysis suggests that while teachers desire a greater level of digitally-mediated assessment literacy to cope in these new environments, they also hold sophisticated, critical perspectives on the technology they must use, and articulate specific needs for more flexible platforms and apps. These early results suggest that effective, technology-driven formative assessment will require a cross-fertilisation of ideas and skills between assessment experts, teachers and educational technology developers.

Reference-Based Selection for Comparative Judgments with Natural Language Processing: Improving Efficiency of Large-Scale Writing Assessments

M. De Vrindt[1, 2], M. Lesterhuis[3], A. Tack[2, 4], W. Van Den Noortgate[1, 2], R. Bouwer[5]

[1]KULeuven, Belgium
[2]Itec, imec research group at KU Leuven, Belgium
[3]UMC Utrecht, Netherlands
[4]KU Leuven, Belgium
[5]Utrecht University, Netherlands

Comparative judgment (CJ) is a reliable and valid method for writing assessments in which assessors repeatedly compare pairs of student works and decide which is of a higher quality. Although using CJ does not require extensive training, it can be time-consuming, as assessors have to make many judgments. This is especially problematic for large-scale writing assessments of essays, since they require a high level of reliability. To make CJ more efficient, it has been proposed to use a reference set of previously assessed essays to which new essays can be compared. Previous research has shown that this reduces the number of judgments. In this presentation, we will show how the efficiency of CJ can be further improved by integrating Natural Language Processing (NLP) into the reference-based comparisons. NLP enables automatic extraction of relevant information from essay texts, which can be used as information for the selection rule in CJ. We will show how such an AI-driven selection algorithm further reduces the number of pairwise judgments needed for reliable text quality scores. We will discuss how integrating NLP for reference-based comparisons provides a valid, reliable, and efficient way of assessing text quality, and what the implications are for large-scale writing assessments.

Development of CVD-accessible assessments as standard

C. Hill-Banks[1], G. Roberts[1], S. Melhuish[1]

[1]WJEC, United Kingdom

This project addresses the significant issue of undiagnosed colour vision deficiency (CVD) affecting up to 8% of male secondary school students and 0.5% of female students (Simunovic, 2010). It is estimated that, of those with CVD, 80% are undiagnosed when starting secondary education (Albany Ward, 2015), highlighting the inadequacy of reactive Access Arrangements. Colour use in assessments can negatively impact performance and confidence for these learners, causing anxiety and confusion. In collaboration with Colour Blind Awareness (CBA), we developed comprehensive guidance and disseminated to over 600 assessment authors and production staff, proving invaluable for teaching practice. This proactive approach ensures assessments are colour accessible at source, negating reliance on additional arrangements. The poster will summarise our research into the issue, our key findings and the work done so far in designing and producing assessments and raising awareness of this issue not only within the context of high-stakes examinations, but in the wider education sector to promote inclusive practices and a universal design mindset, ensuring equality of opportunity from the outset.

References

Simunovic, M. (2010). Colour vision deficiency. Eye, 747-755.

Albany Ward, K. (2015) What do you really know about colour blindness? British Journal of School Nursing, May 2015 Vol10No4

AI-Powered Language Assessment: Can Linguistic and Cognitive Features Predict Item Difficulty and Discrimination?

L. Loxley[1], M. Elliott[1], N. Zanini[2]

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

In many assessment contexts, items are trialled before they are used in live examinations. Within an Item Response Theory (IRT) paradigm, pre-testing is essential to estimate item parameters, i.e. difficulty and discrimination, and ensure item quality. Yet, conventional methods require lengthy and expensive processes/pilots to collect response data from a representative population of prospective candidates, which do not easily/always yield enough observations. The aim of this project is to understand whether, and if so how, it is possible to leverage recent technological advancements to improve item parameter estimation, at pre-testing and in live testing, in the context of English language assessment. More precisely, we are using Natural Language Processing and other AI tools to retrieve the salient features of test items and establish ways to augment the information provided by candidates' responses, thus reducing the sample size needed for live pre-testing. By utilising data from our item banks, we retrieved linguistic and cognitive features of items that can be used to supplement the information provided by candidates' responses. Embedding a novel computational approach to psychometrics within our existing IRT framework will also provide useful insights into our item banks and refine guidelines for item development.

From Selection to Success: Measuring Academic Competencies in High- and Low-Stakes Contexts

M. Sánchez-Martín[1], J.F. Luesia[1], J.F. Plaza[1]

[1]Universidad Loyola Andalucía, Spain

University admission processes offer a valuable opportunity to assess key competencies linked to academic success. However, evaluating these competencies in high-stakes contexts presents challenges, particularly when deciding which instruments are most suitable, especially given the growing emphasis on non-cognitive skills and their susceptibility to faking. This study aims to identify which instruments are most appropriate for use in admissions and best predict academic success.

During the 2024–25 admissions at a private university in Spain, eight tests were administered: four cognitive (numerical, verbal, logical reasoning, and spelling) and four non-cognitive (organizational skills, critical thinking, conscientiousness, and self-efficacy). The same tests were later applied in a low-stakes context during students' second semester.

We have collected over 250 responses and expect to exceed 400 from students across various degree programs. Correlational and regression analyses will be conducted to identify which competencies best predict academic performance, measured via First-Year GPA. We will also explore the impact of faking by comparing scores across high- and low-stakes settings.

Findings will help inform admission procedures by identifying which measures are most effective in capturing a comprehensive view of students' potential for academic achievement.

Supporting the Identification of Learning Needs: Strengthening Science Teachers' Diagnostic Assessment Competence

K. VAINO[1], T. Rosin[1], H. Kann[1]

[1]University of Tartu, Estonia

New knowledge cannot be meaningfully constructed unless relevant prior knowledge is first activated. Yet this essential step is often overlooked, leading to fragmented understanding and persistent misconceptions. Diagnostic assessment plays a key role in identifying what students know, misunderstand, or need next—focusing on instructional needs rather than summative outcomes. Conducted early in instruction, it offers insights into students' prerequisite knowledge and readiness to learn, helping teachers make targeted adjustments.

This poster presents a teacher training initiative aimed at developing science teachers' diagnostic assessment competence, with an emphasis on low-stakes, classroom-based practices such as discussions, probing questions, and student thinking routines. Teachers often focus on assessing lower-order outcomes, limiting their capacity to identify deeper learning needs. The training therefore aimed to broaden teachers' assessment repertoire and shift attention toward diagnosing higher-order thinking and systems understanding. It encouraged a more evidence-based, student-centred mindset, enabling teachers to personalise instruction and pre-empt misconceptions.

To evaluate the impact of the training, changes in teachers' beliefs were studied—not as static traits, but as indicators of evolving understanding and practice. Beliefs were explored through scenario-based responses, and complemented by lesson plans and classroom artefacts.

Results of the pilot study will be introduced at the conference.

Exploring Students' Comprehension of Graphic Texts in the National Assessment of Reading Literacy

T. Hårsaker[1]

[1]University of Oslo, Norway

The purpose of the National Assessment of Reading Literacy in Norway is to assess students' reading literacy and text comprehension as a basic skill in all subjects. The results from the tests are meant to be used as a basis for formative assessment of students learning. The tests are designed from a wide selection of texts within different subjects to measure students' reading comprehension. Some of these texts consists of one or more graphic element. Sometimes visual elements in texts can serve as important carriers of information, even communicating insights that are not present in the verbal part of the text. This poster will focus on how students perform on items connected to graphic texts. By examining students answers on items that are connected to the visual elements in the text, it is possible to gain insight in how students interpret and understand graphic elements. These insights can be helpful for teachers so they can understand what kind of skills and strategies their students may need. Additionally, the findings demonstrate a method by which teachers can find valuable information from students' answers in tests, allowing them to learn what students understand – not just whether they answered correctly or incorrectly.

Real Life – How Much Real Life?

G.A. Nortvedt[1], K.B. Bratting[2], E. Devold[3], H.H. Haram[1], O. Kovpanets[4], A. Pettersen[4], E. Sübül[3]

[1]University of Oslo, Norway
[2]Universitetet i Oslo, Norway
[3]University of Oslo, Department for Teacher Education and School Research, Norway
[4]UiO, Norway

In the Norwegian national numeracy test, students are expected to demonstrate their ability to apply mathematical skills to solve real-world problems. The contexts of the assessment items should reflect applied problems or problem-solving activities relevant to various school subjects, such as social sciences, natural sciences, and interdisciplinary themes like citizenship and sustainability. The assessment development team draws on a wide array of resources, including national-level statistics, newspapers, websites, textbooks, scientific literature aimed at adolescents, and popular books.
However, these resources can often be "wordy," necessitating adaptations to test items in order to prevent excessive reading loads that might introduce irrelevant features. The poster will showcase examples of source texts and the adaptations made to create appropriate test items, illustrating how this process can influence the relevance of the test content and the extent to which the item content can be deemed “real life.”
For instance, simplifying complex data or illustrations—such as those depicting ways to reduce food waste—can reduce a problem to a mere calculation question. This simplification may inadvertently limit test takers' opportunities to express their awareness of environmental issues as well as their ability to apply mathematical skills in real-life contexts.

Assessment Cultures IV

9:00 - 9:30

External moderation of school based assessment: A review of systems' experiences with implementation and factors associated with selecting approaches

D. Murchan[1], S. Shaw[2], E. Likhovtseva[3]

[1]Trinity College Dublin, Ireland
[2]Institute of Education, University College London, United Kingdom
[3]Ravensbourne University, United Kingdom

School Based Assessment (SBA) is increasingly integrated into high-stakes examinations to broaden the range of learning outcomes beyond what traditional tests can capture. However, ensuring consistent and equitable results frequently requires external moderation. This presentation addresses that gap, identifying approaches to external moderation in nine examination systems at the upper secondary level and exploring the factors influencing their adoption and implementation. A two-phase sequential survey design was employed. Phase 1 involved review of publicly available information from 23 examination systems, identifying nine illustrative systems for further examination. Phase 2 involved interviews with senior officials, exploring the rationale and lived experiences of implementing externally moderated SBA. Six themes emerged as shaping policymaking: curriculum and learning outcomes; quality assurance; fairness and equity; teacher professional development; resource availability; and degree of central authority oversight. Challenges were also revealed such as teacher scepticism towards statistical mark adjustments, inconsistent assessment expertise, and substantial resource limitations. Combining policy-oriented and practical insights, this study provides enhanced understanding of how externally moderated SBA can be tailored to diverse educational contexts. Findings highlight the importance of transparent communication, robust teacher support, and sustained investment in moderation capacity to enhance validity, fairness, and feasibility in high-stakes assessment systems.

Implementing an Assessment Framework in Cox's Bazar refugee camps: two years on

G. Billings[1], A. Oomeer[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 have continued their work with UNICEF to develop and implement an assessment framework for children being educated in the camps, alongside a training programme and a baseline and endline evaluation of assessment practice.
This presentation will focus on the enactment journey, which has included extensive engagement within the vulnerable Rohingya community to design and implement an assessment framework that meets the needs of stakeholders and the reality of an astonishingly difficult environment. Challenges along the way have included managing stakeholder engagement, working with incomplete curricular, consistent assessment in an emergency context, upskilling of teachers, translanguaging, and high rates of community illiteracy.
Including authentic voices from numerous observations and focus groups, this paper aims to empower the community we are representing, as well as providing an interesting overview of implementing assessment in an emergency refugee context.

Bridging perception and performance in teacher appraisal: a multi-faceted study from Kazakhstan

N. Shora[1], B. Yessingeldinov[1], A. Shilibekova[1], A. Zhumykbayeva[1]

[1]National Center for Professional Development "Orleu", Kazakhstan

This study examines the teacher appraisal (certification) process in Kazakhstan, with particular emphasis on its validity, reliability, and impact on professional development. It investigates whether teachers' perceptions of the appraisal system align with their actual instructional performance in the classroom. Earlier analysis identified ten underlying constructs, three of which varied significantly across qualification categories: teachers' perceptions of the teaching methodology section within the certification assessment, the appraisal system overall, and use of regulatory documents.

To validate these perception-based findings, a classroom observation instrument was introduced to compare teacher self-assessments and external instructor evaluations across key instructional dimensions, analysed in relation to qualification categories as indicators of professional growth.

A multi-method quantitative approach was applied, using more than 100,000 teacher assessment records, survey responses from over 600 teachers, and classroom observation data from more than 20,000 teachers and 8,000 instructors. Observations focused on planning, instruction, student interaction, assessment (monitoring), and reflection. Analyses in R-software included descriptive statistics, factor analysis, and reliability testing.

Applying Kane's argument-based approach to validity, the study integrates multiple sources of evidence to assess alignment between perceptions and classroom performance. Findings support a more transparent, evidence-informed teacher appraisal model that reflects both self-reported and observed professional competence.

Assessment of Practical Skills III

9:00 - 9:30

Borderline Regression Analysis: A Data-Driven Framework for Fairer Educational Decisions

D.T. Kropmans[1]

[1]Qpercom Ltd, Ireland

Abstract
Borderline Regression Analysis (BRA) has emerged as a powerful, evidence-based method for standard setting in clinical assessments such as OSCEs. Unlike arbitrary pass marks, BRA uses the relationship between checklist scores and examiner global ratings to determine fair, defensible cut scores for student performance. This approach enhances decision accuracy, reduces subjectivity, and aligns with best practices in psychometric evaluation. By integrating BRA with the Standard Error of Measurement (SEM), institutions can minimize false pass/fail outcomes and support transparency in high-stakes assessments. Research demonstrates BRA's effectiveness in improving reliability, examiner consistency, and actionable feedback for learners. Its applications extend beyond undergraduate education to include medical and dental recruitment, notably in NHS England's selection processes. Here, BRA contributes to equitable candidate evaluation and data-driven decision-making. As digital platforms like Qpercom incorporate BRA into real-time analytics and feedback systems, the method's impact is further amplified. In a landscape demanding fairness and accountability, BRA offers a robust framework for standard setting that supports both educational integrity and learner development.
References:
Milan et al., 2022; Park et al., 2021; McGowenet al., 2022.

Implications of different types of test items on applicants' results in school-leaving exams

S. MIrzayeva[1], N. Abdurahmanova[2]

[1]the State Examination Center of the Republic of Azerbaijan, Azerbaijan
[2]The State Examination Center of the Republic of Azerbaijan, Azerbaijan

This study investigates the impact of different types of test items on applicants' results in school-leaving examinations in Azerbaijan. In 2019, open-ended and coded-response items were introduced alongside traditional formats to better assess students' skills, knowledge, and application abilities. The research examines how these changes in assessment methods influenced applicants' overall performance and the effectiveness of the national examination system. Using a quantitative, retrospective comparative design, the study analyzes secondary data from the official examination database, applying the Analysis of Variance (ANOVA) method via R Studio software. The sample includes data from eight secondary schools with consistently stable performance across six examination years: 2012, 2015, 2018, 2019, 2021, and 2024. Results show that exam year is a statistically significant factor affecting average student scores, suggesting that variations in test item types contributed to differences in achievement levels. The findings emphasize the critical role of test design in ensuring valid and fair assessments and highlight the importance of aligning assessment formats with educational objectives. The study provides insights for improving exam frameworks and supporting evidence-based educational policy reforms in contexts involving high-stakes national examinations.

The challenge of designing an inclusive, high-stakes, digital baseline assessment for 4-year-olds in England

M. Hornett[1], C. Whitby[1]

[1]Department for Education, United Kingdom

From 2028, primary school progress in England will be calculated using a new baseline measure utilising data gathered from 4-year-olds. This has necessitated the development of a new, statutory, high-stakes assessment delivered to young pupils within their first 6 weeks at school.
2025 marks the introduction of a new digital version, designed to be delivered on a 1:1 basis using paired devices.
Early research helped to shape item types and response methods, and content domains were developed to link with the curriculum pupils follow throughout their future schooling.
We will outline some of the challenges we encountered in developing an inclusive and accessible assessment for all learners, including designing appropriate accessibility settings and access arrangements to ensure fairness and equity for all.
The delivery of the assessment is reliant on the existing IT infrastructure within schools in England. This adds further demand to successful national roll-out.
In this session, we will share:
• insights gained from user research during development
• observations from the launch of the updated assessment
• feedback received from stakeholders
We will also explain how we aligned Agile user-centred design with our extensive trialling model to develop a bespoke assessment that is valid, reliable and fair.

Psychometrics and Test Development III

9:00 - 9:30

Exploring the impact of optimising the consistency of examiner scores used in model training on automarker performance.

T. Breakspear[1], H. Bouteba[1], M. Brenchley[1], A. Chakravarty[1], I. Lewin[1], T. Sajimon[1], Y. Huang[1]

[1]Cambridge University Press & Assessment, United Kingdom

Whilst rightly emphasising the importance of data quality in automated essay scoring systems (AES), language assessment practitioners have tended to focus their attention on the quality of evaluation data. This study shows how practitioners can contribute to enhancing AES performance by optimising training data, presenting a novel yet easily implementable approach to optimising the consistency of examiner scores. Specifically, we incorporate a distance metric, Earth Mover's Distance (EMD), to create a subset of examiners based on the proximity between their individual scoring distributions and the overarching distribution of the population dataset, as sampled from a high-stakes, English language writing assessment. We create a randomly sampled control set with the same distribution and sample size as the optimised set. We train both classical, feature-based models and deep learning, LLM-based models on the control and optimised sets and evaluate them on an all-mark-all set of 1692 responses marked by 15 examiners.
Although the feature-based model did not benefit from the optimised data, the LLM-based model outperformed control across multiple metrics, including a 50% improvement in EMD. We conclude that optimising training data primarily using a distance metric has the potential to enhance LLM-based AES particularly in replicating the desired distribution of scores.

Targeting the top: Evaluating differentiation among high-performing candidates in items designed to stretch and challenge

A. Ulicheva[1], S. Hughes[1, 2]

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

This study examines the effectiveness of two research item types - Respond to a Situation and Summarise Group Discussion - in a high-stakes language proficiency test. Designed to be more cognitively demanding, these items aim to better differentiate high-performing candidates by targeting advanced language skills. We evaluate whether they achieve this goal using secondary analysis of large-scale field test data.

Our approach focuses on item performance patterns, comparing two matched groups: (A) top performers who excel on the research items and (B) individuals who perform similarly on standard items but do not score as highly on the research item types. By contrasting data sets with and without the research items, we assess differentiation using classical test theory and psychometric analysis, as well as descriptive statistics and entropy measures (e.g., Shannon entropy as a measure of uncertainty).

We also examine score profiles across traits to determine whether group A demonstrates stronger performance in areas linked to higher-order skills. To further explore proficiency differences, a linguistic analysis of responses investigates whether group A produces more complex language on the research items.

Findings offer empirical evidence on the discriminatory power of these item types and inform test development for assessing advanced language proficiency.

Inclusive Assessment I

9:00 - 9:30

Redesigning assessments for inclusion: A developmentally-aligned, criterion-related approach for students aged 11-16

S. Chakrabarty[1], S.M. Easwardas[1], L. Dutton[1]

[1]International Baccalaureate, Netherlands

This paper presents the redesign of International Baccalaureate (IB) Middle Years Programme (MYP) assessment model, focusing on a developmentally-aligned, criterion-related approach aimed at promoting greater inclusivity and equity. While the MYP has long emphasized student-centred learning and authentic assessment, challenges persist in implementing its principles across diverse contexts shaped by norm-referenced grading systems and national assessment frameworks. The model's perceived rigidity can appear misaligned with the flexibility required to support students with learning needs, language acquisition challenges, or non-linear academic trajectories, especially in high-stakes assessment contexts.

The redesigned assessment model introduces a two-phase structure – lower phase MYP (for students aged 11-14) and upper phase MYP (for students aged 14 -16) - to better reflect cognitive and emotional development and to scaffold expectations accordingly. Drawing on findings from relevant educational research and the Enhancing the MYP: Assessment trial, this paper examines the redesigned assessment model, reduce misalignment with national systems, and support learner diversity more effectively. Particular attention is given to educator and learner perceptions of fairness, clarity, and usability in the redesigned model, especially in multilingual and inclusive classrooms.
It offers insights into how international assessment frameworks can balance rigor, flexibility, and equity in diverse educational environments.

The Way You Teach Matters! How Different Teaching Practices to Develop Reading Comprehension Skills Impact Italian Students' Reading Abilities

S. Mazzuca[1]

[1]INVALSI, Italy

This study investigates the relationship between teaching practices and reading comprehension skills among 6,000 Italian students and their 400 teachers, utilizing data from PIRLS 2021. Employing SEM, the research identified three key latent factors that characterize the diverse teaching strategies prevalent in Italian classrooms. The subsequent analysis reveals that the implementation of specific instructional approaches exerts a statistically significant influence, effectively differentiating the reading performance of students across the nation. These identified factors encompass various aspects of reading instruction, including explicit strategy teaching, engagement promotion, and differentiated instruction – tailoring instruction to meet diverse learner needs (readiness, interest, learning profile).
The findings highlight the substantial impact that deliberate pedagogical choices have on students' reading comprehension outcomes, establishing the critical role of teachers' instructional decisions. The presentation will delve into a detailed examination of these three pivotal teaching factors, elucidating the specific strategies encompassed within each and highlighting their differential influence on student reading achievement. The research aims to inform strategies that enhance reading instruction and improve literacy outcomes for all students nationwide. The implications emphasize the necessity of adopting targeted and responsive teaching practices to effectively address the inherent diversity within educational settings and foster strong reading abilities.

Enhancing inclusion and cross-cultural validity in PISA. Towards an improved index of economic, social and cultural status

F. Avvisati[1], S. Ettejjari[1], A. Osses Vargas[2]

[1]Organisation for Economic Co-operation and Development, France
[2]Australian Council for Educational Research, Australia

Significant efforts are undertaken in PISA for developing contextual questionnaires to collect data on students' attitudes, home environments, and school learning conditions. This helps to contextualise student achievement and to monitor educational equity.
A key construct for examining educational equity across PISA participants is students' economic, social, and cultural status (ESCS). However, concerns have been raised about the reliability and validity of the PISA ESCS index for reflecting the realities of participants while still allowing for international and over time comparisons.
For PISA 2025, the OECD introduced several changes to improve ESCS measurement. For the PISA field trial, a new list of items was proposed for the home possessions index, allowing countries to select contextually relevant items for their national questionnaire. New items measuring living conditions and material deprivation were added. The wording of items collecting information about parental education and occupation was modified to reflect modern family structures.
In this study we discuss how these innovations contribute to designing inclusive contextual instruments that better reflect the realities of students participating in PISA and allow for reliable cross-cultural comparisons. Data from the PISA 2025 field trial is used to support these claims.

Formative Assessment II

9:00 - 9:30

Making Computational Thinking visible as framework for inquiry and assessment

T. Knight[1], R. Roberton[1]

[1]International Baccalaureate, Netherlands

Computational thinking (CT) made visible triggers an inquiry process that supports pedagogical knowledge for mathematics with implications for summative and formative assessment. The heuristic inquiry process using CT is discussed in the context of problem solving, mathematical modelling and statistical thinking. CT as a framework (Figure 1) supports an inquiry based curriculum, identification of assessment objectives for summative assessment, encouraging a pedagogy of dialogue in an inclusive classroom and embedded formative assessment.

Figure 1: DP Mathematics: Subject Framework

The framework also informs the design of the assessment model. The core represents the purpose of mathematics, to engage in inquiry and problem specification. The central circle can be assessed in coursework, whilst the three sections of the second circle can be observed in formal, written examination contexts. The inner two circles are used to develop assessment objectives for three levels of demand, which in turn will inform examination papers, mark schemes and criteria type marking guidance. The outer circle houses the essential skills and attributes required to successfully deploy the inquiry process, but cannot easily be observed in isolation. The need to observe these in interactions highlights the role of formative assessment in learning.

The effects of the launch of the Learning Outcomes Framework on teaching, learning and assessment in the French as Foreign Language Classroom. The case of Malta.

R. Bonello[1]

[1]University of Malta, Malta

The introduction of the Learning Outcomes Framework (LOF) (2015) brought about a series of changes that impacted teaching, learning and assessment practices in the French as Foreign Language (FFL) classroom. In fact, the design and implementation of the syllabi based on the LOF in secondary schools as well as the introduction of School-Based Assessment (SBA) to weigh in the national certification system operated by the MATSEC Examination Board of the University of Malta are changes aimed at favouring the link between learning and assessment.
In this paper, the local documentation published by the Education Department and MATSEC is examined to highlight the rationale behind the assessment reform. To find out the impact of the new syllabus design and assessment system on learning, a selection of interactions between teachers and students from a corpus of 48 lessons collected in 4 secondary schools are analysed in the light of the theory supporting formative assessment and the principles underpinning the LOF. Data obtained through a focus group interview with collaborating teachers and a questionnaire distributed to teachers of French in Malta underlines the implications of the new assessment system in the FFL classroom and teachers' understanding of assessment in relation to learning.

Artificial Intelligence and Assessment IV

9:00 - 9:30

Towards Scalable Personalised Feedback: Technology-Supported Insights into Student Feedback Use and Self-Regulated Learning

F. van der Kleij[1], A. Lipnevich[2], T. Hopfenbeck[3, 4]

[1]Australian Council for Educational Research, Australia
[2]City University of New York, United States
[3]University of Melbourne, Australia
[4]Norwegian University of Science and Technology, Norway

Personalised feedback is widely recognised as a key mechanism for improving student learning, yet providing personalised feedback at scale remains a challenge. Advances in educational technology, including artificial intelligence, create new opportunities for more timely and personalised feedback. However, the impact of feedback on learning is highly variable, shaped by how students perceive, interpret, and act on it within their self-regulated learning (SRL) processes. The existing research base lacks sufficient evidence to support targeted personalisation of feedback. To address this critical gap, the paper (1) presents a new conceptual model that explicitly links feedback use and SRL, highlighting the reciprocal relationship between external feedback and students' affective, cognitive, and behavioural responses, and (2) proposes methodological approaches to examine how students engage with feedback in technology-enhanced learning environments. We argue for a shift from retrospective self-report measures to technology-enabled micro-level methods that capture in-the-moment feedback use and SRL processes. Specifically, we introduce a micro-analytic Feedback Self-Regulated Learning (F-SRL) protocol designed to examine how students engage with feedback before, during, and after assessment tasks. This approach offers a foundation for future research and practical applications, generating insights into how feedback can be optimised to support self-regulated learning and personalised feedback at scale.

ENHANCING AUTOMATED ESSAY SCORING WITH ARTIFICIAL INTELLIGENCE (AI): THE ROLE OF NATURAL LANGUAGE PROCESSING (NLP)-BASED ASSESSMENT TOOLS IN IMPROVING SCORING SYSTEMS

O. OYENIYI[1], R. OJO-ODIDE[1]

[1]WAEC, Nigeria, Nigeria

The Assessment of essay tests has long been a challenging task for educators. The issues of subjectivity, inconsistencies and fatigue have always plagued traditional essay evaluation methods. Artificial Intelligence (AI)-driven Automated Essay Scoring (AES) systems have emerged as a promising solution to these challenges, providing a more reliable and consistent scoring process. This paper examines the current landscape of AI-driven AES by analyzing the core functionalities of Natural Language Processing (NLP)-based AES tools, highlighting their role in improving scoring accuracy, standardization, and feedback mechanisms. The research adopts a mixed-method design, using quantitative analysis for scoring accuracy and qualitative analysis for system usability. The methodology involves training AI models on a large essays dataset using NLP models and doing a comparative analysis of the system against human raters to assess reliability and bias. Findings revealed that NLP-enhanced AES improved consistency, fairness, and scalability of essay evaluation within a shortened grading time. Despite these advancements, biases in training data and challenges related to contextual understanding remain serious considerations. Recommendations include continuous refinement of AI models using diverse and representative datasets, hybrid approaches combining AI with human supervision to ensure fairness, and policy interventions to support the adoption of AES in educational assessments

The future of non-examined assessment (NEA) – malpractice detection and the challenges and opportunities of generative AI

T. Leech[1, 2], R. Laugalyte[1, 2]

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

Non-examined assessment (NEA) is any assessment not taken under controlled exam conditions, including what is commonly known as coursework. The rise of generative AI in recent years poses potential threats to NEA validity, if a) it is possible for candidates to produce responses using such technology, b) such use contravenes regulations for that assessment and c) if this usage is undetectable. In this presentation we will review evidence about how learners are engaging with AI, whether legitimately or illegitimately. Drawing on a multi-part programme of research, we will then look at what we know so far about how well illegitimate use of AI can be detected, either by software systems such as Turnitin's AI detection or by people, using methods such as comparison to examined work or looking for AI hallmarks. Finally, in dialogue with wider literature, including from higher education contexts, and thinking more broadly about detectability, innovative assessment design and what constructs and skills we want NEA to assess, we will explore what these findings could imply for future NEA approaches.

Fairness and Social Justice II

9:00 - 9:30

Towards the automated optimisation of examination timetables

A. Furlong[1], D. Den Hertog[2], J. Grand Clement[3]

[1]International Baccalaureate, Netherlands
[2]University of Amsterdam, Netherlands
[3]HEC Paris, France

Every exam session the International Baccalaureate (IB) faces a significant challenge in creating an examination timetable for more than 60 exams within a three-and-a-half-week window that must adhere to various rules such as consecutive days for exams within the same subject and a maximum of six hours of exams per day. Currently, the timetable is compiled manually. This study explores the use of mixed-integer linear optimization to automate the creation of an optimal examination timetable with minimal manual input. The study first defines a 'good' timetable, considering factors like student stress caused the proximity of pairs of exams to one another. This definition, along with the IB's timetabling rules are then formulated as a mixed-integer linear optimization model. Although the model could not be fully solved within a reasonable time using the full IB dataset, it produced a valid solution within a reasonable time. Improvements to the model, such as forcing exams within a subject to be in a given order, showed promise in reducing computation time. The study concludes that while further development is needed before use in a live IB examination session, the model offers a promising and viable approach to making exam scheduling fairer and easier.

Strengthening Assessment and Learning through Culturally and Linguistically Responsive Design

L. Badham[1], C. Constantinescu[2]

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

Linguistic and cultural diversity amongst cohorts raises critical questions of fairness and validity in assessment design. Linguistic and cultural elements influence how exam questions are interpreted, and what skills are valued. For example, assessments often exhibit a 'monolingual bias', favouring monolingual students and neglecting multilingual skills. This can create negative washback, such as teachers discouraging translanguaging in the classroom, fearing it may hinder performance on high-stakes assessments. As diverse cohorts become the norm, alternative approaches are needed to ensure fair and valid assessments that support inclusive teaching and learning.

This presentation discusses a case study of an internationally recognized "language and culture" qualification designed for linguistically and culturally diverse students. The pilot qualification includes a linguistic autoethnographic assessment task. Drawing on document analysis, focus groups, and survey data, the study explores the culturally and linguistically responsive design of the task.

Findings suggest the task was perceived as “empowering” and “validating” for multilingual students, but still posed challenges for those with lower English proficiency. It highlighted tensions between reliability and construct validity in designing for diverse cohorts. Nevertheless, it exemplifies a socioculturally responsive design, that affirms multilingual learners’ skills and experiences, marking a shift away from a monolingual bias in assessment.

Keynote Speech

11:00 - 11:45

Raising educational standards

J. Baird[1]

[1]University of Oxford, United Kingdom

A perennial challenge for the field is how assessments can support learning and thereby raise educational standards. Part of the answer to this must involve clear messaging about what the educational standards are, to provide goals for teachers and learners.

In this address, I will consider how standards are thought about differently in three assessment paradigms: construct-, curriculum- and outcomes-based approaches. I have previously characterised these paradigms as prototypical ways of thinking about assessments, since many assessment systems are hybrid versions, borrowing across paradigms. In each case, the philosophy of assessment differs in terms of the attribute of interest, the definition of standards and expectations of outcomes. This creates different relations between what kind of thing is valuable learning, the ways in which we think they ought to be assessed and which quality criteria we prioritise. Different paradigms promise information of different kinds about standards for teachers and learners.

In research conducted with over 900 stakeholders in Scotland during the pandemic, we investigated how stakeholders think of national assessments and their standards. One of our main findings was that they think of the same assessments in relation to different paradigms. I will outline how this raised significant issues for the management and communication of standards by exam boards, that go beyond misunderstandings or lack of assessment literacy. In a smaller-scale qualitative study in Wales, we investigated how industry-insiders and teachers thought about standards and their communication. A striking finding from this study was that not only were teachers unaware of the standard setting processes, but they did not believe that they needed to know. This raises questions about exactly what teachers and learners need to know about standards and whether our industry-insider perspective has been miscommunicating.

e-Assessment IV

11:45 - 12:15

Face Validity in Focus: Engaging Test-Takers in the Development of High-Stakes English Language Tests

E. Barrow[1]

[1]Pearson Education, United Kingdom

This session will focus on a study exploring test-taker perspectives and experiences of potential revisions to a global high-stakes English language test. The primary aim of this study was to gather insights into test-taker experiences of two research item types requiring extended spoken responses, scored by a combination of human and AI judgement, that reflect real-world language use. By examining these perspectives, we aim to evaluate the face validity of these items to ensure they meet the needs and expectations of their intended purposes. We also explored secondary themes to elicit a more rounded understanding of the overarching test experience, test preparation practices and most notably, attitudes to the use of AI scoring in high-stakes language testing. Feedback on the items indicated that these tasks were perceived to be relevant and reflective of real-life scenarios, thereby supporting the test's face validity. Moreover, the findings showed a general trust in AI scoring systems. However, there is a consensus on the need for a balanced approach, combining AI and human oversight to ensure accuracy and reliability. Overall, the research underscores the significance of incorporating test taker feedback in the test development process to maintain a robust and trustworthy assessment tool.

Automating Effective Feedback for Comparative Judgment Assessments of Essays

M. De Vrindt[1, 2], T. van Daal[3], M. Lesterhuis[4], D. Mortier[3], W. Van Den Noortgate[2, 5], A. Tack[2, 5], R. Bouwer[6]

[1]KULeuven, Belgium
[2]Itec, imec research group at KU Leuven, Belgium
[3]University of Antwerp, Belgium
[4]UMC Utrecht, Netherlands
[5]KU Leuven, Belgium
[6]Utrecht University, Belgium

Comparative Judgment (CJ) is an assessment method in which assessors make multiple pairwise comparisons of essays. While quality scores from CJ are both reliable and valid, they lack transparency regarding how assessors make their judgments. Assessors can provide feedback comments, but this is time-intensive.

Our research addresses these feedback limitations of CJ by developing automated feedback for CJ using Natural Language Processing (NLP). We study whether more effective feedback can be produced for CJ using NLP. To automate feedback, we predicted different aspects of writing quality from essay texts to explain the quality scores from CJ. Subsequently, we developed an interactive dashboard showing both automated and human feedback. The feedback is presented in a comparative way that allows students to compare their essay and feedback with that of higher-quality essays. This way, students can learn how to improve their essays.

We conducted a mixed-methods study in which students used our feedback dashboard to improve their scientific writing. Data was collected through surveys, log data, and interviews. Our findings show that the automated feedback complements assessors' feedback. Students readily accepted this feedback, as they reported that they intended to use it to improve their essays.

National Tests and Examinations III

11:45 - 12:15

Evidence-Based Prompting: Searching for the Holy GrAIl

O. Hrubes[1]

[1]Scio, Czech Republic

This study presents a three-phase methodological framework for identifying qualitative parameters of high-stakes test items with superior discriminatory power in order to define the ultimate prompt for test items generation. The results demonstrate the benefits and limitations of current Large language models within the field of Social Sciences high-stakes testing.

The discrimination index, a psychometric measure quantifying an item's ability to distinguish high-performing from low-performing examinees, serves as the cornerstone in the process of high-stakes tests production. As per classical test theory, items with DI values ≥0.35 demonstrate optimal discriminatory efficacy. Usually it proves to be very difficult to estimate DI of test items without a dry run with student specimens, especially in the field of Social Sciences. But what if we knew exactly which parameters make the test items DI optimal? Could we then generate the whole tests with the help of LLM using this information?

The high-stakes test items generation using LLM brings major challenges. It proved useful to make a thorough 3 phase examination before utilizing the benefits of LLM for effective and high-standard quality assurance of high-stakes test items in Social Sciences. Nevertheless the human factor is still hugely needed in the whole process and iteration.

Beyond oracy – 30 years of using Oral assessments to assess candidates' knowledge in high stakes exams at 16 and 18.

A. Evans[1], J. Maziarz[1]

[1]WJEC, United Kingdom

Since 2001 there have been over 1,200 different GCSEs (General Certificate of Secondary Education for 16-year-olds) and over 1,300 different GCEs (General Certificate of Education for 18-year-olds) approved by regulators in England, Wales and Northern Ireland. Whilst oral assessments have been used to assess fluency in a second language and presentation skills they have not been used as an alternative to a written paper for other assessment objectives. The exception has been oral assessments of Welsh Literature which have taken place in high stakes qualifications since 1995.

This analysis is a quantitative and qualitative look at how these assessments have performed. Quantitatively, the correlations between the oral performance and written items assessing the same assessment objectives are examined. Qualitatively, interviews with moderators and assessors have been undertaken looking at the perceived reliability and validity of these assessments.

The discussion then looks at the envisioned English Language and Literature GCSE in Wales which adopts a similar form of assessment and whether this method of assessment could be expanded to other qualifications to provide alternative forms of assessment in a world where AI is limiting options for controlled assessment.

Scottish educators' attitudes towards AI use in the classroom: A mixed methods consultation

M. Mroczkowski[1], J. Green[1], J. Lawson[1], R. Whitford[1]

[1]SQA, United Kingdom

The Scottish Qualifications Authority (SQA) conducted a two-phase consultation with educators in Scotland to gain an understanding of their views around and uses of AI in education and assessment. A survey was administered (phase 1) followed by in depth focus group interviews (phase 2) with schoolteachers, college lecturers, employers and training providers and other education stakeholders. We present the findings from these two phases of research.
A total of 519 educators completed the online survey. Participants had a range of prior experiences with AI, including for material or activity design. Participants emphasised that SQA must prioritise AI detection and cultivate the skills necessary to critically evaluate AI generated sources and use the technology responsibly. Similar themes emerged from follow up focus group interviews with 33 participants, which probed further into the potential implications of AI in education and what support was needed. Participants called for clear, sub-specific guidance in order to help learners adapt to and engage with the new landscapes offered by AI technology ethically and effectively. This research has helped to inform SQA’s guidance around the use of AI in its centres and has also helped to shape the next stage of our consultation work involving learners’ views.

Assessment Cultures V

11:45 - 12:15

Dialogic Assessment Conversations as Dynamic and Learning-Oriented Assessment in English Language Education

E.L. Hauer[1]

[1]University of Inland Norway, Norway

This study explores how English teachers in Norwegian lower secondary schools use dialogic assessment conversations (DACs) as a tool for formative and dynamic assessment within the framework of learning-oriented assessment (LOA). Rooted in sociocultural theory, DACs are structured teacher-student dialogues that promote learning by integrating feedback, reflection, and scaffolding into the assessment process. The project investigates how these conversations support inclusive education by accommodating diverse student needs and fostering metacognitive awareness, student agency, and language development. Using a qualitative approach, the research draws on semi-structured focus group interviews with teachers, video-stimulated recall interviews, classroom observations, and student logbooks. Hermeneutic phenomenology was applied to teacher interviews, while thematic analysis was used for video recall data, and student experiences were explored through case study methodology. Findings indicate that DACs, when used as dynamic assessment, can shift the assessment focus from grading to growth, supporting both language learning and student motivation. Teachers play a key role as mediators, using real-time dialogue to tailor instruction and assessment. The study contributes to the understanding of how LOA principles can be practically applied through DACs, offering insights for policy development, teacher education, and classroom practice in language assessment.

Are political beliefs associated with assessment policies?

M. Meadows[1], J. Baird[2]

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

AEA-Europe has fostered better understanding of the different national contexts in which assessments are understood. Research on assessment has recognised the historical, societal and institutional factors that shape these understandings. The sociology of assessment has also drawn attention to the ways in which assessment can be used as a technology by politicians, to shape society in neoliberal ways, for example. We also observe the position that assessment, and education more broadly, should be taken out of politics. However, due to their society-shaping capacity, education and assessment are highly political matters, even when accepted as the status quo. With shifting geopolitics, matters which seemed settled are more open to question. As such, empirical research investigating individuals' political views and whether there is an association with their assessment policy preferences is timely. Although many in our field are involved in policymaking or influencing, the political dynamics are often obscured. Having a better understanding of these associations would illuminate policy positions. This presentation reports on interviews with policy elites (politicians, policymakers and influencers) in England of different political persuasions, to investigate the extent to which individual's political beliefs predict their assessment policy preferences.

Reframing Dialogues of Learning, Assessment and Mastery: Insights from Norwegian Lower Secondary Schools

E.W. Hartberg[1], K. Blichfeldt[1], I. Jacobsen[1], K. Haaland[1], T.S. Wille[2]

[1]Faculty of Education, Inland Norway University, Norway
[2]The Education Agency, Oslo Municipality, Norway

This paper builds on findings from the Dialogues on Learning and Assessment (DOLA) study, including DOLA1 and DOLA2 (2024), which examine how formative assessment is implemented in line with the 2020 revision of Norwegian national assessment regulations. Despite policy emphasis on motivation, agency, and daily formative practices, both early career and experienced teachers report difficulty reconciling these aims with the dominance of summative logics. Students often view formative feedback as a predictor of final grades, reflecting broader international concerns about high-stakes assessment and its procedural impact on classroom practice.

In response, a new research phase focuses on transformative assessment dialogue. Through action research, we collaborate with schools, teacher networks, and local authorities to reframe practices through shared reflection and experimentation. Focus group interviews in 2025 will further explore evolving teacher perspectives on formative versus final assessment.

Preliminary findings show a strong desire among teachers to develop practices that strengthen assessment as a learning dialogue. Promising examples, such as co-authored success criteria and metacognitive tools, emerge in schools with well supported professional learning communities. This paper addresses how formative assessment, when dialogical and co-constructed, can better align with curriculum intentions and support student learning, motivation, and mastery.

Artificial Intelligence and Assessment V

11:45 - 12:15

AI in Assessment: So much more than ChatGPT

N. Thompson[1], T. Trierweiler[1]

[1]Assessment Systems, United States

Artificial intelligence (AI) has received an astounding amount of attention since the release of ChatGPT in November 2024. Most of the discussions regarding the use of AI for assessment have revolved around automated item generation, since ChatGPT is so good at producing written content. However, the applications of AI in the world of assessment is far greater.

First, we will discuss the use of large language models (LLMs) to generate test questions, item review, item categorization, simulated pilot testing, or essay scoring. We will also discuss the concerns for this approach, including the fact that it is a black box and therefore does not provide support for validity, intellectual property law, and security concerns.

We will then discuss non-LLM applications of AI to assessment, including automated essay scoring, adaptive/multistage testing, online proctoring, and adaptive learning. These approaches are typically done with custom-calibrated machine learning algorithms, which requires more effort and expertise but also then provides sound validity documentation rather than being a black box.

Finally, we will present some real-world examples of these methods, including supporting research and lessons learned. We then discuss approaches for responsible AI use with assessment.

Abilities aren't numbers: Alfred Binet and the road not taken in educational assessment

A. Scharaschkin[1]

[1]AQA, United Kingdom

Representing assessment constructs as numerical quantities is deeply embedded in educational assessment. From their origins in psychological testing, quantitative latent variable methods have become ubiquitous in the practice of educational assessment.

Yet Alfred Binet, the originator of IQ tests, did not regard the constructs that his tests assessed as being quantitatively structured. He conceived of intellectual capability as being better represented as (in Joel Michell's words) an 'ordered attribute with heterogeneous differences between its degrees', reflecting the idea that 'the cognitive states underlying
test performance are not quantitatively homogeneous, but differ from one another in heterogeneous ways'.

Building on Binet's conception, this talk will discuss a non-quantitative mathematical approach to discriminating between, and ordering, learners, with respect to their proficiency in an educational domain. It represents assessment constructs as hierarchies of cognitive configurations, or states, that are qualitatively discriminable from each other with respect to construct-relevant attributes.

The talk will use school examination data from England to show how this approach provides a new way of appraising examination grading standards over time, and will also briefly discuss implications for the application of large language models to grading students' qualitative responses to assessment tasks.

Evaluating AI-Generated Feedback: Insights on usefulness, accuracy, relevance and actionability

L. Howarth[1], K. Gilmartin[1]

[1]AQA, United Kingdom

This research investigated the potential for AI to enhance student learning and support teachers in the classroom.
An initial study focused on evaluating whether AI-generated feedback can be produced at a level suitable for students to use and understand. For three different subjects, we provided AI specialists with a training dataset of approximately 1,000 student responses to train their models. Subsequently, they were given 24 hours to generate feedback for a test dataset comprising around 100 unseen student responses for each subject. Expert examiners reviewed a sample of the AI-generated feedback for each subject and participated in questionnaires and interviews to evaluate the quality.
A follow-on study trialled AI-generated feedback on short English essays in a classroom setting, aiming to explore AI's potential to deliver rapid, detailed and personalised feedback. The feedback platform was trialled with students aged 15 (n = 111) and teachers (n = 4). The results highlighted the importance of specifying the types of question best suited to AI-generated feedback.
Both studies highlighted teachers' recognition of AI's potential to support students and teachers. However, appropriate use of AI is crucial to avoid undermining the student-teacher relationship and hindering student learning.

Comparative Judgement II

11:45 - 12:15

Graded Response Modeling of Forced-Choice Measures

D. DIMITROV[1], D. ATANASOV[2]

[1]Education & Training Evaluation Commission, Saudi Arabia
[2]New Bulgarian University, Bulgaria

This paper presents a new approach to scoring of multidimension forced-choice (MFC) questionnaires where the items in each MFC block measure different traits. The response data is organized in the traditional MFC format of pairwise scores for item preferences based on the Thurstone law of comparative judgments. Trait scores are presented on a bounded scale using graded response modeling in the framework of the D-scoring method of measurement (DSM). The proposed method, referred to as graded response modeling of forced choices (GRMFC), avoids the complexity in design and estimation procedures of extant MFC methods such as the popular Thurstonian IRT (TIRT) method. Using simulated data, it is shown that the trait scores obtained via the GRMFC and TIRT methods are highly correlated. Furthermore, the GRMFC method has some advantages in estimation accuracy, avoids issues of model fit and convergence associated with TIRT, and provides information about cumulative and category-specific probabilities of endorsing MFC items, as well as a statistic for detection of response “faking.” A syntax code (in R) for GRM-D computations is also provided.

Investigating the validity of comparative judgement: What influences judges' decisions when assessing STEM subjects?

M. Walter[1]

[1]Ofqual, United Kingdom

Further research evidence is required to demonstrate the validity of using comparative judgement to determine assessment performance, particularly when the assessment to be evaluated consists of a large number of items as is typical in STEM subjects. This study investigated judges' decision-making when using comparative judgement to assess performance in highly structured multi-item STEM exams. Expert judges were asked to evaluate students' performances in GCSE and A level assessments in five STEM subjects by making a series of pairwise comparisons. Concurrent (think-aloud) and retrospective (survey) self-reporting methods were used to collect data relating to judges' decision-making processes when forming their judgements. The results of this analysis reveal several key features of the assessed work that judges focused on as indicators of student ability, as well as judges' attentiveness to several construct-irrelevant features. Additionally, this study identified some arguably undesirable judge behaviours that may lead to construct underrepresentation and could therefore represent a threat to validity. These behaviours are attributed, in part, to the high task complexity and intrinsic cognitive load associated with making holistic comparisons across a large number of individual assessment items, compared to the relatively more straightforward evaluation and comparison of a single pair of extended tasks.

Beyond Reliability: Challenges of Using Adaptive Comparative Judgment in High-Stakes Legal Assessment

K. Egelandsdal[1], J. Færstad[1]

[1]University of Bergen, Norway

The study investigates the use of Adaptive Comparative Judgment (ACJ) as an alternative to traditional criterion-based grading in two third-year law courses at a Norwegian university. Nine examiners assessed 107 exam papers by making holistic, pairwise judgments of anonymized responses. In one course, examiners provided short written justifications for each decision, while in the other they did not. Median judgment times ranged from under a minute to over 20 minutes, reflecting this design difference. While many exam papers received similar outcomes across ACJ and traditional grading, some cases showed notable divergence, with ACJ decisions appearing to emphasize clarity and structure over deep legal analysis when decisions were made rapidly. The length and complexity of the exam papers may have amplified this issue, as rapid judgments risk overlooking deeper qualities in extensive responses. Although ACJ allowed for quick individual judgments, the overall workload remained substantial due to the large number of required comparisons. Examiners also highlighted challenges related to transparency and fairness, particularly the difficulty of providing individualized, criterion-based feedback in a high-stakes legal education context. The study suggests that while ACJ has potential for supporting assessment consistency, careful attention must be paid to validity, workload, and requirements for grade justification.

Test Development

11:45 - 12:15

The impact of response format on performance in a maths test for gifted and talented primary children

E. Darlington[1]

[1]Cambridge University Press & Assessment, United Kingdom

This study investigates the impact of response formats on student performance in a mathematical reasoning test typically used to identify gifted and talented primary age children for selective education.

Two parallel trial tests were developed that were in a similar style to questions used in live mathematical reasoning tests. The 35-question tests were taken on an online platform familiar to participants. Test A employed a standard 5-option multiple choice format for every question. Test B asked the same questions, but with 29 questions using alternative response types. This included open-ended, multi-select and drag-and-drop ordering response formats.

Self-selecting participants the same age as the target candidature of the live tests completed the tests in exam conditions during the school day. The online platform recorded scores for each item, and the time participants took on each question.

The research aimed to determine whether response formats had a significant impact on performance and completion time. The results offer an insight into the use of varying response types in assessments of this type, particularly when considering the challenges associated with writing good assessment questions, compiling fair and balanced tests, and identifying those who are gifted and talented in mathematical reasoning.

Key principles for determining the optimal item format

E. Sweiry[1]

[1]Ofqual, United Kingdom

Despite ample literature addressing the strengths and limitations of selected response (SR) and constructed response (CR) items, there is limited guidance to support practitioners in determining optimal item format in a given context. Additionally, there is a scarcity of research or guidance on the utility of different SR item formats. The paper presents a set of principles intended to support practitioners in determining the optimal item format, based on features and properties of the content and skills they wish to assess. The principles were informed by a literature review and empirical study. In the empirical study, biology, psychology and business item writers 'converted' CR items from past papers into a range of SR formats, assessing the same knowledge and skills as the original items as far as possible. An additional group of examiners reviewed each item pair, using a coding frame to make judgements about the impact of converting from CR to SR on properties such as language demands, item difficulty and task clarity. Assessment design decisions are the subject of complex balances and trade-offs, and the choice of item format is a key factor within these decisions.

Achievement Levels in Irish Literacy among Post-Primary Students Educated in Irish-Medium Schools

M. Bergin[1], H. Ní Rócháin[1], C. Ó Duibhir[1], B. Donohue[1]

[1]Educational Research Centre, Ireland

This paper focuses on achievement levels in Irish literacy among post-primary students educated through the medium of Irish, by examining the results of the standardisation study of MDLI-G (Measúnú Diagnóisic Litearthachta d'Iarbhunscoileanna – Gaeilge), an Irish-language screening test designed to assess specific literacy skills of post-primary students in grades 7 and 9. It explores how various factors, including gender, school type, home language, and language of primary-level education, are associated with literacy outcomes in Irish. It also seeks to investigate how students' literacy levels in Irish compare to their performance in English, specifically by comparing the results of MDLI-G with those of PPAD-E (Post-Primary Assessment & Diagnostic – English), the English-language equivalent test.

Other

11:45 - 12:15

Redesigning assessment objectives for authentic and aligned learning and assessment: moving away from Bloom’s taxonomy

S. Sorrentino[1]

[1]International Baccalaureate, Netherlands

The presentation focuses on the re-design of the assessment objectives in the International Baccalaureate (IB) Diploma Programme (DP) Visual Arts guide published in 2025. The redesign improved integration of curriculum and assessment, one of the main goals of the review. Embedding the assessment objectives within the language of the subject practice and the creative process, together with the improved structure of the summative assessment enhanced the experience of the student as an art practitioner and the value of art-making as inquiry. The assessment model remains rigorous and in line with IB assessment principles while allowing much more space for authentic student experiences. Inspired by the work of Eliott W. Eisner, the new DP visual arts guide moves away from ‘instructional’ objectives – using the Bloom’s taxonomy verbs - to a framework adopting what Eisner defines “expressive objectives”. This change supports a more integrated and truly creative curriculum where the learning of each student is not only evaluated but also valued.

The presentation will introduce collaborative iterative curriculum design process and how the use of visuals supported the development as well as the presentation of the new integrated framework leading to more aligned and authentic learning and assessment.

Decolonising Assessment in South African Education: Towards an Ubuntu-Centered Approach

A. Kanjee[1]

[1]Tshwane University of Technology, South Africa

Drawing on the intellectual work emanating from the 'decolonise education movement', this paper analyses the impact of the Assessment for Learning (AfL) approach adopted by South Africa's Department of Education to incorporate principles of Ubuntu so as to address the learning needs of all learners.
The paper begins by discussing the role of the "decolonising education" movement in creating new spaces to challenge current colonial-based systems that promote performativity and test-based accountability regimes. It then analyses the impact of two key initiatives aimed at providing alternative policies and practices that foreground the lived experiences of learners and teachers. First, the curriculum strengthening process implemented to promote AfL as a pedagogical strategy in all schools; and second, the adoption of alternative capacity development programmes based on African philosophy of Ubuntu (I am because you are) that not only addressed specific learning needs of all key role-players but also the unequal power relations that continues to serve as an obstacle to effective policy implementation.
The paper concludes by highlighting key AfL policy and practice successes and discussing challenges and lessons for the schooling system in South Africa as well as other countries embarking on similar change initiatives.

An adaptive national assessment: technical and policy challenges and insights

A. Boyle[1], C. Hope[2]

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

This paper concerns a national computer adaptive test (CAT), whose objective is to promote formative assessment. This 'formative CAT' exists in an environment where there are concerns about learners' achievement. Assessment sponsors see this national system as the natural source of attainment information. We discuss analyses we undertook to provide a longitudinal national report, including an exposition of the technical challenges involved in generating a meaningful item response theory (IRT) calibration. On the one hand, we want to use the most up-to-date and accurate calibration, but we also need to manage any changes to previously published reports that updated calibrations could give rise to.
We also discuss how best such a national system can provide information about learners' attainment by the end of primary school. In doing so, we observe:
• Spread of scores is as important as average attainment levels.
• No set of 'quintessential items' that can be a proxy of end of primary demand.
• We need to unpack the notion of 'Year 6 questions' – deciding whether this holds water, empirically and conceptually.
Using a CAT for these purposes brings particular challenges. But the challenges that adaptivity brings are always present when assessments provide attainment information.

Artificial Intelligence and Assessment VI

14:15 - 14:45

Improving OSCE Feedback Using ChatGPT-4.0

D.T. Kropmans[1], K. Mahrlamova[2], A. Alsahafi[3]

[1]Qpercom Ltd, Ireland
[2]University of Dundee, United Kingdom
[3]University of Galway, Afghanistan

Feedback in Objective Structured Clinical Examinations (OSCEs) is essential for developing clinical competence. Yet, examiner feedback is often inconsistent, generic, or delayed, undermining its educational value. Studies at Dundee Medical School and in University of Galway show that global rating scores often diverge from checklist scores, and written comments lack the depth students need for meaningful improvement.

This study evaluated ChatGPT-4.0 as an automated feedback tool. Using anonymized OSCE data—scores, global ratings, and examiner comments—ChatGPT generated structured, personalized feedback aligned with learning objectives. A mixed-methods approach assessed its efficiency, clarity, and actionability.

ChatGPT consistently produced well-structured feedback faster than human examiners. It reduced variability and emphasized learning outcomes, though it occasionally lacked context sensitivity. The most effective approach combined ChatGPT’s speed and structure with human review to address nuance.

In conclusion, ChatGPT-4.0 offers a scalable way to enhance feedback quality in OSCEs. While not a complete replacement for human insight, it supports more timely, specific, and educationally meaningful feedback. A hybrid AI-human model presents a promising direction for modernizing assessment feedback in medical education, where continuous validation research is required.

Let me know if you'd like this version included in your slide deck or printed material.

AI-Driven Predictions of Mathematics and Science GCSE Exam Results using Mock Papers

S. Nastuta[1]

[1]Pearson Education, United Kingdom

This paper presents a novel approach to accurately predict students' future exam performance, thereby providing valuable assistance for schools in preparing for high-stakes examinations. It evaluates the extent to which and the accuracy with which future results can be predicted using mock results, a common formative practice in the UK.
Utilising a supervised machine learning approach, item-level performance data from GCSE Mathematics and Science examinations held in June 2023 were employed to predict 2024 exam outcomes for students who sat as mock exam papers shadowing the 2023 exams.
Several machine learning algorithms, such as OLS regression, logistic regression, support vector machine, decision tree, and K-Nearest Neighbours (KNN), were employed to train and develop several mark prediction models using results from all candidates who sat the GCSE exams in June 2023. Early in 2024, some students took shadow papers of the 2023 exams as mock exams. Their performance in this formative assessment was employed to predict their future performance in the 2024 examinations. Comparing the predicted grades with the actual grades for the matched candidates, we concluded that different machine learning algorithms provide between 80% to 95% accurate predictions, making this approach valuable for schools.

Incorporating structural elements in Automated Question-Difficulty Estimation models can improve model accuracy and efficiency

G. Ortega[1], A. Jiménez[1], S. Lions[1, 2], P. Dartnell[1, 3]

[1]Centro de Investigación Avanzada en Educación (CIAE), Instituto de Estudios Avanzados en Educación (IE), Universidad de Chile, Chile

[2]Departamento de Evaluación, Medición y Registro Educacional (DEMRE), Universidad de Chile, Chile

[3]Centro de Modelamiento Matemático (CMM) y Departamento de Ingeniería Matemática (DIM), Universidad de Chile, Chile

Automated Question-Difficulty Estimation (AQDE) holds significant potential for educational assessment by surpassing human difficulty estimates, cutting pre-testing costs, and boosting the effectiveness of online applications such as Computerized Adaptive Testing (CAT).

Several models have been proposed that distinguish between structural elements of multiple-choice questions, such as stem, key, and distractors.

However, there is a lack of studies comparing the impact of these structural elements on model performance, particularly regarding the role of option order. This paper evaluates the effect of incorporating different sets of structural elements on AQDE model error.

Our models achieved R^2 metrics comparable to those reported in the state-of-the-art, with values of 0.83 and 0.71 for science and history questions. The highest accuracy was observed when the distractor encoder was restricted to a single shared encoder rather than assigning a separate encoder to each distractor. Furthermore, making the output of the distractor encoder order-invariant maintained a similar level of accuracy while substantially reducing the number of model parameters.

These findings suggest incorporating structural elements can help develop scalable AQDE models suitable for large-scale assessment. The best model's reduced parameter count makes running these models locally feasible, which has important implications for privacy and security in test development.

Assessment Cultures VI

14:15 - 14:45

Time, the one-handed clock, and educational certification theory

R. Harry[1]

[1]WJEC, United Kingdom

This research proposes an extension of Newton et al's (2004) nascent 'educational certification theory' by examining how temporal constraints and manageability concerns shape qualification design processes. Using Wales' recent development of new GCSEs as a case study, I analyse the tensions that emerge when curriculum principles meet traditional assessment structures, particularly where curriculum requirements are determined first in order to frame qualification needs and purposes, rather than the two being designed in conjunction. The qualification design process in Wales was further complicated by its distributed nature across multiple agencies and shifting priorities over time. Drawing on Stobart's (2008) "one-handed clock" concept, I demonstrate how manageability considerations and teachers' resource constraints are a core element of anticipatory qualification design that need to be made explicit in a theoretical framework, and show how purposes and proficiency modelling can be defined so that tensions at trade-offs can be made drawn out and resolved more coherently even as new concerns arise. This research contributes to assessment theory by highlighting the dynamic interplay between idealised design principles and the pragmatic realities of educational systems undergoing reform.

The importance of political and cultural embeddedness in the enactment of Assessment for Learning: Lessons from the “Network of schools for Assessment Innovation (Red sin Notas)” in Chile

M.T. Florez Petour[1]

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

Given the evidence on the positive effects of formative assessment on student learning, the question rises as to why research also shows difficulties for a more widespread enactment of the approach in classrooms. This paper is aimed at offering lessons from a highly unique case of a research-practice partnership known as the “Network of schools for Assessment Innovation (Red sin Notas)” in Chile, a context where the enactment of AfL is particularly complex, due to the predominance of an assessment culture focused on high-stakes testing and summative assessment. This paper presents the results of a strand of work of the Network that involves the collection of evidence on its development. Two rounds of data collection and analysis were conducted within the network, consisting of focus groups with teachers and school leaders. Findings offer insights on postive aspects and challenges in the processes transformation of assessment cultures in the Network, and conclusions highlight the relevance of considering the political context and the assessment culture in which the enactment of AfL is embedded, as well as the obstacles that are experienced in the process, as constitutive aspects of professional development initiatives.

Formative Assessment III

14:15 - 14:45

Exploring the Formative Potential of Standardized Reading Tests: Enhancing students' competence in critical reading through assessment results

C. Weyergang[1], T.S. Frønes[1]

[1]University of Oslo, Norway

This study examines the formative potential of standardized reading tests and their role in fostering student learning, particularly in critical reading. Globally, standardized reading assessments serve various purposes, such as informing policymaking and guiding classroom instruction. There is an increasing emphasis on enhancing the formative aspects of these tests, a discussion that has also gained traction in Norway, where national reading assessments have been implemented since 2004.

Formative assessments are designed to provide feedback that can modify teaching practices. Effective assessment relies on several key components: the construct must be clearly defined, and appropriate items must be developed to yield evidence of students' competencies. This study explores how the construct of "critical reading" can be incorporated into standardized tests, and how teachers can use the results to improve instructional practices. Critical reading entails engaging with texts reflectively, which is vital for informed citizenship.

Drawing on examples and data from a national test for 5th graders, we illustrate how test results can enhance instructional practices. We emphasize that analyzing incorrect responses can uncover students' misunderstandings, thereby facilitating targeted classroom activities aimed at improving critical reading skills.

Formative Assessment with Impact: What can we learn from different contexts

B. Wyatt[1]

[1]Cambridge University Press and Assessment, United Kingdom

The definition and practice of formative assessment varies across different countries. This paper explores how formative assessment is understood, applied, and supported in different contexts, and considers how these differences affect the impact on student learning.
Drawing on qualitative data from different education systems, the paper examines how perception, policy, and teacher professional development all influence the implementation of formative assessment. While the intent to improve learning using formative assessment is supported by all stakeholders in different contexts, practices differ due to conflicting factors and limitations.
Findings suggest that the successful use of formative assessment depends on how well it is embedded in the wider system of curriculum and pedagogy and not just on policy. Factors such as teacher confidence, clarity of purpose, and use of assessment information all influence whether formative assessment leads to meaningful learning.
This paper contributes to the wider picture of assessment reform and provides practical insights for policymakers, school leaders, and teachers.

National Tests and Examinations IV

14:15 - 14:45

Exploring the use of similar items for monitoring grading standards

C. Draper[1]

[1]Ofqual, United Kingdom

In England, grading standards are set using statistical evidence and expert judgement. An assumption of this is that, if the cohort has not markedly changed from one year to the next, it is unlikely that the outcomes should change markedly either. It is important to explore the extent to which this assumption holds and how it may interact with genuine changes being reflected in grades. Common items across multiple assessment series can indicate changes in performance, as their difficulty remains constant. Due to assessment security risks in high-stakes assessments like GCSEs and A levels, the analysis of similar items—defined as eliciting comparable performance to previous questions—is suggested instead. This approach could provide evidence to better identify changes in performance over time. The study focused on A level and GCSE maths, involving two awarding organisations (AOs) with large entries. Five senior examiners per AO reviewed 2023 and 2024 exam papers to identify similar items, considering content, format, and cognitive demand. The study established criteria for final item selection and highlighted features of items which made the task difficult. The results inform us of the potential to use similar items as evidence for the maintenance and/or monitoring of standards.

Raising educational standards in the Kingdom of Bhutan: a case study of an assessment framework

J. Frith[1], G. Billings[2]

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

In 2024, a needs analysis was carried out in Bhutan at the invitation of the Ministry for Education to identify key actions towards achieving the aim of aligning their curriculum to international standards regarding competences and the application of knowledge. As well as a desk review of key curriculum and assessment documents, this involved extensive engagement with officials, teachers and students. The findings were synthesised into a report which included a number of recommendations.
Following this, an extensive programme of work has been agreed upon to support these objectives. This involves a multitude of activities including curriculum alignment, training of key ministry personnel, developing test specifications, and improving assessment practices and instruments. The aim of this is to eventually allow benchmarking of Bhutan's Grade 12 leaving qualification with international standards and co-certification.
This presentation will focus on the context of the project within the national education aims of Bhutan and will focus primarily on our approach to developing a coherent assessment framework. It will also discuss some of the key challenges faced in the work so far, including public perception, international standards alignment, recognition, ambitious timelines, and translanguaging.

The Predictive Validity of High-Stakes Testing: A Meta-Analysis of Long-Term Academic and Career Outcomes

R. Wilson[1], B. Owusu-Ansah[1]

[1]The West African Examinations Council, Ghana

High-stakes standardized testing has long been a cornerstone of educational assessment, serving as a gatekeeper for college admissions, professional certifications, and policy decisions. Despite widespread use, debates persist about its efficacy in predicting long-term success, with critics arguing that such tests fail to capture essential non-cognitive skills. This study synthesizes existing research to evaluate the predictive validity of high-stakes testing on academic and career outcomes, addressing gaps in the fragmented literature. The study would employ meta-analysis of longitudinal studies. Eligible studies to be used will tracked participants for a minimum of five years post-assessment, control for socioeconomic variables, and provide measurable effect sizes. The study would compare high-stakes testing against alternative assessment methods, such as continuous evaluation and competency-based models, to determine relative predictive strength.

Preliminary findings suggest that high-stakes testing demonstrates meaningful correlations with later academic achievement and early-career success.

Based on the evidence, the study recommends retaining but refining high-stakes testing frameworks to address its shortcomings to ensure both rigor and fairness in educational evaluation.

Fairness and Social Justice III

14:15 - 14:45

From Deficit to Asset: Reframing Validity Through Linguistic Diversity in Classroom-Based Assessment

P. Nagpal[1]

[1]Center for Educational Measurement & Assessment , University of Sydney, Australia

In multilingual classrooms, students from minority language backgrounds are often labelled as underperforming, despite assessments being designed with limited regard for their linguistic repertoires. This conceptual paper challenges the deficit narratives that dominate assessment discourse in such contexts and proposes an asset-based reframing of validity in classroom-based assessment (CBA), emphasizing the strengths and potential of all students rather than their perceived shortcomings.
Drawing on Solano-Flores' matrix of evidence for validity argumentation, this paper conceptualizes validity as a dynamic, context-sensitive construct shaped by the intersection of language, culture, and pedagogy. It asserts that conventional assessment practices often overlook how students' home languages, dialects, and linguistic resources influence their task interpretation, knowledge expression, and feedback engagement.
Using field-based examples from Indian classrooms, the paper outlines a framework that positions students' multilingualism as a resource, not a barrier. It introduces three interrelated validity dimensions for equitable classroom assessment: linguistic accessibility, cultural relevance, and instructional alignment.
By shifting from deficit-based to asset-based perspectives, this paper encourages educators, researchers, and policymakers to reconceptualize validity not merely as a psychometric criterion but as a means of promoting justice, fostering meaningful learning, and affirming the linguistic identities of all students.

Minimising construct-irrelevant barriers for neurodivergent learners in assessment: Exploring neuro-inclusivity by design

R. Chivers[1], L. Badham[2]

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

Construct-irrelevant barriers must be minimized in assessment to ensure fairness and validity. However, there is limited understanding of how assessment design processes can effectively dismantle barriers for neurodivergent learners. Increasing emphasis is placed on embedding universal design principles early in assessment design to better support neurodivergent learners. However, further research into neuro-inclusive assessment design is needed.

This presentation discusses a scoping review that explored assessment design in the context of neurodiversity. The aim was to develop evidence-based recommendations for neuro-inclusive assessment design by removing construct-irrelevant barriers for neurodivergent students. Following PRSIMA guidelines, the review ensured broad coverage of relevant peer-reviewed literature related to inclusive assessment design for neurodivergence. Overall, 23 sources were selected, data was methodically charted, and a grounded theory approach was taken for data analysis.

Despite limitations in representing diverse neurodivergent profiles in the literature, findings reveal a shift towards a social model of inclusive assessment design. This entails a holistic approach, identifying characteristics typical under the neurodivergent umbrella rather than categorizing them by individual neurodevelopmental differences. Assessment design recommendations are then made based on these characteristics. Such approaches may ultimately, result in more valid assessments that better reflect all students' skills and abilities.

High Stakes and High Accountability: Public Post-Test Item Review in Czech University Admission Process

B. Lunakova[1]

[1]Scio, Czech Republic

At Scio, we create and administer a wide range of standardised tests which are used as university entrance exams by one third of Czech university faculties. As part of our effort to foster accountability, transparency, and fairness, we invite test-takers to publicly challenge specific items on their tests in a secure online environment after each test administration. These item appeals may concern correctness, ambiguity, or fairness of individual test items. The test development team prepares formal responses to each such appeal, which are then reviewed by an independent review committee composed of subject experts. The committee's decisions determine whether items are upheld or removed, or if two correct answers are accepted due to item ambiguity. Our company adheres to the committee's decisions in the scoring process. The present paper aims to introduce this process of Public Item Review at Scio, as well as encourage discussion about similar endeavours and good practice regarding item review with experts from other countries, contexts, and assessment cultures, with the common goal of testing fairly, transparently, and with a positive impact on stakeholders.

Summative Assessment

14:15 - 14:45

Enhancing science teachers' assessment literacy through authentic modelling of multiple assessment methods

B.S. Haug[1], S.M. Mork[1]

[1]University of Oslo, Norway

In this study, we explore a multiple summative assessment (MSA) approach with various types of exams during one-year professional development programs (PDP) for Norwegian secondary science teachers. This approach challenges traditional summative assessments and offers students ample opportunities to demonstrate their knowledge and competencies in various ways. Assessment methods were modelled, and the different types of assessments served as exams to pass the PDP, thus teachers experienced various assessments in an authentic setting. Our research question "How may authentic exposure to various summative assessment methods influence teachers' perspectives on learning and assessment?" was investigated based on 37 reflection notes and 8 individual interviews regarding teachers' views and experiences of the assessments. Findings indicate teachers favoring alternative assessments over traditional ones, linking this to learning outcome and opportunities to demonstrate competencies in various ways. The authentic setting raised teachers' awareness of their own assessment practice, especially the importance of explicit learning goals and assessment criteria. Challenges included students' stress in submitting many exams, and teachers' workload providing feedback and grades. This suggests that MSA needs to be balanced between contextual aspects and opportunities for students to demonstrate their competencies.

Going gradeless in upper secondary school. A multiple case study of a natural experiment

H. Fjørtoft[1], L.V. Sandvik[2], I. Kohanová[1], S. Tveit[3], D. Normann[4], A. Amdal[1], S.A. Angvik[1], A. Seljeseth[1], K. Smith[1]

[1]NTNU Norwegian University of Science and Technology, Norway
[2]NTNU, Norway
[3]University of Agder, Norway
[4]Norwegian university of science and technology (NTNU), Norway

Grades are central to most education systems but are often associated with negative effects on student wellbeing, motivation, and teacher workload. In response, some schools are adopting “going gradeless” (GG) approaches to foster deeper learning and reduce performance pressure. This study investigates how GG is implemented and experienced in seven Norwegian upper secondary schools.

Using a multiple case study design, we conducted group interviews with teachers, students, school leaders, and parents. Our analysis revealed five key findings: (1) substantial variation in how GG was introduced, supported, and evaluated; (2) growth in teacher assessment literacy over time; (3) a heightened focus on feedback and communication in teacher-student relationships; (4) mixed student responses, with both reduced and increased stress reported; and (5) strong alignment between GG practices and Norway’s high-trust, student-centered education culture.

The findings suggest that GG may be an important component in assessment innovation but involves significant cultural shifts in schools. However, this change demands careful implementation, sustained professional support, and alignment with broader school values and community expectations. Specifically, GG requires careful consideration of the role of grades, feedback, and other ways of communicating with stakeholders about student learning processes and outcomes.

Higher Education and Assessment

14:15 - 14:45

Understanding the Access Arrangements System in Wales

A. Harrison[1]

[1]Qualifications Wales, United Kingdom

Qualifications Wales, the regulator of non-degree qualifications in Wales, requires all awarding bodies to comply with The Equality Act 2010 by making available reasonable adjustments and access arrangements. However, we do not stipulate the details of what arrangements should be in place. As a result, we set out to conduct research to address the following research questions: How is the access arrangements system designed, including by awarding bodies not part of the Joint Council for Qualifications? How does the access arrangements system function in practice, both at the awarding body level and at the centre level? We analysed a range of awarding body documentation before conducting interviews with awarding body participants. We then visited different centre types and interviewed Additional Leaning Need Coordinators, Exam Officers and learners. By engaging with stakeholders working in different contexts, the research provides a holistic overview of the system as well as insight into both contextual and wider systemic challenges. Importantly, the study is not without limitations, including the inability to generalise findings given the research design. It is intended that the findings of this study will inform the lines of enquiry for further research in this area.

To Test or Not to Test. Evolving Landscape of Higher Education Admissions in the Czech Republic and Slovakia.

A. KUTARNA[1]

[1]Scio, Czech Republic

This paper examines the evolving landscape of higher education in the Czech Republic and Slovakia, focusing on the challenges faced by universities, particularly regarding admission exams. It highlights the significance of these exams, often considered more critical than the high-school leaving certificate (“maturita”), and explores the attitudes of higher education decision-makers in the context of the COVID-19 pandemic and the emergence of Artificial Intelligence. Will this be considered as another threat leading to omitting testing altogether or will the insitutions strive for a different focus in their examinations?
By analyzing statistical data and incorporating insights from discussions with various stakeholders, the study aims to identify prevailing patterns of thought among academic leaders and predict how potential students may respond to different admission policies when selecting their fields of study. The findings contribute to a deeper understanding of the current educational climate and the implications for future university admissions in the region.

Symposium Session 1 - Integrating assessment and curriculum design: harnessing analysis of TIMSS data to positively impact the enacted curriculum.

16:15 - 16:45

Understanding TIMSS data: an insider critique of the frameworks, assessments and questionnaires

G. Grima[1]

[1]Pearson UK, United Kingdom

Paper 1 focuses on the development of the assessment frameworks for mathematics and science, the domains assessed in this international assessment together with the context questionnaires completed by pupils, teachers and school principals. It explains the development of the frameworks including the updating process at the start of each cycle and discusses the involvement of key stakeholders which include expert groups and participating countries. It decribes the adaptation and translation processes of the assessments and questionnaires at national level and then goes on to discuss the alignment with national curricula of participating countries. Countries also have the opportunity to add items of national interest in the questionnaires. The context questionnaires in this ILSA contribute significantly to our understanding of the assessment results, and their usefulness for informing developments in the enacted curriculum. Additionally, TIMSS published an encyclopedia ahead of the international results which presents a profile of the national contexts shaping mathematics and science education in each of the countries that participated in TIMSS2023. Through information collected from the TIMSS2023 curriculum questionnaires as well as country-authored chapters, the encyclopaedia serves as an important vehicle for comparing and contrasting the common and unique features of the country contexts and curricular goals used in teaching and learning mathematics and science around the world.

TIMSS 2023 evidencing mathematics and science education as widely gendered endeavours: the issues and some pedagogic (and broader) responses

J. Golding[1]

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

Globally, participation and attainment in mathematics and science are widely, though not universally, gendered – often in favour of boys. In the twenty-first century this raises issues for equitable personal and societal thriving. ILSAs are important since they can offer low-stakes, authentic and longitudinal evidence of persistent learning, and attitudes/experiences – in the case of TIMSS in both mathematics and science in grades 4 and 8 as well as an international comparative and longitudinal lens.

This paper analyses the widespread increases in gendered patterns of TIMSS performance and questionnaire responses between 2019 and 2023, including in grade 4; the related issues appear particularly acute in England. TIMSS questionnaires probe pupils' reported experiences with, and attitudes towards, mathematics and science as well as broader education; these are supplemented by contextual data from (all or some of) pupils' teachers, Headteachers and parents.

In response to the scale of identified issues, we argue a need to draw on curriculum, pedagogy and teacher education literature to both understand emergent patterns and reverse-engineer the related curriuclum policy and practice. We synthesise at a high level approaches evidenced to be productive in addressing the exposed gendered inequities. In common with much available data, we adopt a (simplified) binary conceptualisation of gender.

We don’t belong here: views on school belonging absenteeism and achievement

M. Richardson[1]

[1]UCL Institute of Education, United Kingdom

Regular school attendance is recognised as a key factor in determining pupil achievement and engagement with education. In England, since 2019, persistent absenteeism - pupils missing at least seven or eight days of school per term - has increased. The 2023 Trends in Mathematics and Science Studies (TIMSS) questionnaire responses revealed similar trends globally across many participating countries and jurisdictions. Reasons for persistent absenteeism are complex and can include factors such as socio-economic status, parental influence and health/well-being. Choosing to be absent from school suggests a disaffection with, or disconnection from, education and absentee pupils may not feel they ’belong’ in school. It appears that the Covid pandemic and its impact globally, has negatively impacted some pupils’ connectedness with education and the TIMSS participant survey responses revealed some negative attitudes towards learning, attendance, and a varied sense of belonging in school.
This paper presents further analysis of these themes and explores the factors related to persistent patterns of absence during and since the pandemic. Using TIMSS data, we will present arguments for the development of curriculum initiatives for enhancing pupil belonging in schools to demonstrate that value of leveraging such data to better understand and address societal challenges and their impact on education.

Symposium Session 2 - Assessing higher order thinking skills in an Assessment for Learning perspective : Challenges and prospects facing teachers in Switzerland

16:15 - 16:45

Under which conditions implementing criterion-based assessment ?

R. Pasquini[1]

[1]University of teacher education Vaud, Switzerland

In the 7 French speaking cantons, there are wide disparities in terms of the importance given to AfL in each canton policies. Some documents are very explicit, others are less. In the canton of Vaud, which is the biggest one, policies are obviously embedded in broader federal school orientations. However, it's one of the few cantons where AfL is theorized including summative assessment. Therefore, in line with this contemporary

conceptualization, many prescriptions and law articles highlight the fact that assessment has to be criterion-based. On one hand, this approach represents an opportunity for teachers to develop assessment practices that can better support students' learning. On the other hand, criterion-based assessment remains one of the biggest challenges teachers face. Thus, in the light of a study case lead at the primary school by a teacher, we will describe the conditions under which this approach of assessment can become a powerfull tool for supporting student's learning and improve teachers' decisions. Moreover, we will identify several concerns and issues in a critical perspective, acknowledging that contextual and subject factors play a decisive role in this kind of practice shift.

Does a new curriculum require new assessment tools?

R. Smit[1]

[1]St.Gallen University of Teacher Education, Switzerland

With the introduction of the new Curriculum 21 for the German speaking part of Switzerland, more complex and application-oriented abilities and skills are to be examined at the end of three cycles. These three cycles each span multiple K-12 grades. In addition, the teacher should also support the acquisition of these more complex skills with formative assessment. Most cantons asked their school administration to develop guidelines on this topic and recommend assessment tools such as portfolios or rubrics, which are not yet used very frequently in the classrooms.

In general, the construction of rubrics is considered to be rather time consuming and difficult by the teachers. Furthermore, the question arises as to the quality with which such rubrics are created, given that the descriptions of the criteria are demanding. While rubrics can be used independently of pedagogical practices, portfolios require a higher level of independence on the part of students, which not all teachers consider useful or feasible. The question of whether the possibilities of such instruments for the increased use of formative assessment are seen is also open. In our presentation we will take a closer look at the question of assessing higher-order skills based on these descriptions.

How can teachers implement a competency-based assessment embedded in a backward planning approach ?

M. Salvisberg[1, 2]

[1]University of Teacher Education (DFA/ASP), Switzerland
[2]University of Applied Sciences and Arts of Southern Switzerland (SUPSI), Switzerland

Teachers in the Canton of Ticino are facing a period of cultural transformation with regard to assessment: the new version of their Curriculum published in 2023 follows a competency-based approach. Jointly, a book on assessment for teachers was also published. This document, aligned with the curriculum, is based on the following key concepts: Assessment for Learning, assessment for competences, backward planning and assessment rubrics. Currently, at the cantonal level, continuous training courses are organized to support teachers' assessment practices shifts. In this contribution, we will highlight two significant changes of this new approach for teachers: (1) modifying teaching/learning and assessment practices rooted in objective-based pedagogy towards a competency-based approach. For example, through assessment rubrics, the aim is to keep track of learning progression, focusing on processes (higher skills); (2) designing assessment assignment considering backward planning and using rubrics. Through interviews with teachers about the design of their assessment process developed during various training sessions, we will identify and discuss challenges and prospects teachers face within the two significant shifts mentioned above.

Symposium Session 3 - Capturing Diverse Student Voices in Assessment

16:15 - 16:45

Study on the education and well-being of disabled pupils in France

E. Persem[1]

[1]French Ministry of Education, DEPP, France

In France, students with disabilities are most often educated in mainstream schools. The introduction of the law of 11 February 2005 on Equal Rights and Opportunities affirmed the principle of the inclusion and educability of all pupils by establishing the right of disabled students to be educated in the school closest to their home. In October 2013, the Depp (Direction de l'évaluation et de la performance, Department of Evaluation, Forecasting and Performance) set up a panel of disabled pupils born in 2005 in order to find out about the educational pathways of these children in particular, as well as their academic level in mathematics and French. The last survey took place in 2021, at the end of compulsory education. A conative questionnaire asked the students about their well-being during their schooling. This included their hopes for their educational outcomes, their expectations for their future careers, their motivation, their anxiety and how they had experienced the Covid 19 pandemic in terms of their schooling. The results of this study are informing the public debate on the challenge of educating disabled students and the challenges and opportunities of school inclusion.

Trialling a New Numeracy Product: Capturing students' experiences of maths learning in Further Education and Alternative Provision settings

S. Miller[1], F. Walker[1]

[1]AQA, United Kingdom

This project employed an inclusive methodology to explore the responses to a prototype numeracy assessment of students underrepresented in research, namely those resitting lower secondary exams in GCSE mathematics and students studying in Alternative Provision (AP) settings, such as hospital schools. Seven focus groups were carried out with students (n=24) in their educational settings. Questions explored students' views of the prototype assessment, challenges they had faced studying maths and whether they felt the content they had studied was applicable to their everyday life. To ensure inclusivity, we responded to specific requests from AP providers, modifying timings to accommodate medical treatment and establishing a predictable pattern of 'turn-taking' during focus groups for students with autism. A reflexive thematic analysis of the data was undertaken. Many students reported previous negative school experiences and attitudes in maths. Students recognised the importance of numeracy as a key skill, but felt that much of the current GCSE maths content lacked relevance to everyday life, valuing the real-world relevance of the prototype assessment. 'Gamification' features rewarding progress and a clear, predictable interface were key to engagement. Accessibility and customisation of the interface were also important to students. This work will shape the development of a new approach to learning and assessment of everyday numeracy skills that can meet the needs of diverse learners.

Multilingual students' performance in high school history classes: teaching and assessment needs

T. Rousoulioti[1]

[1]Aristotle University of Thessaloniki, Afghanistan

A system of education that includes all the students and respects children's rights encourages them to learn regardless of who they are, what they can do, or what their requirements are (Unicef, 2017). The continuous increase of plurilingual students in high schools and the need to accommodate their teaching and assessment needs across the curriculum was the motivation behind this research study that investigated how plurilingual students are assessed in school subjects other than language. The current study was undertaken within the realm of the history course in a Greek public intercultural high school in Athens, Greece. The research was conducted using both a qualitative (interviews with the 6 teachers) and a quantitative research design (questionnaire completed by 42 plurilingual students). The results showed that the limited knowledge of the second language (L2) - mainly in academic vocabulary - contributes to a lack of understanding of the academic language in history textbooks and corresponding exam questions, resulting in low performance of plurilingual students. In addition, the assessment processes cause students to feel insecure, anxious, and indifferent at times. Finally, both teachers and students expressed a preference for using digital translation applications and implementing teamwork and alternative assessment methods in history courses.

Symposium Session 4 - Rating systems for dynamic assessment in adaptive learning environments

16:15 - 16:45

Tracking progress in subdomains: When are multiple ratings better than one?

A. Kuchina[1]

[1]Tilburg university, Netherlands

In adaptive learning environments, accurately tracking student progress is essential for providing effective feedback. Originally developed for chess, the Elo rating system has been adapted as a dynamic assessment method for student abilities. This study investigates whether skill domains should be split into subdomains, each with a separate Elo rating, or treated as a unified domain with a single rating. Although splitting domains can improve diagnostic precision and item selection, it also reduces the amount of data per subdomain, potentially requiring larger updates to ability estimates after each response, which can lead to noisier measurements. Furthermore, the benefit of using subdomain ratings may depend on the strength of correlations between the subskills.
Building on previous research in educational testing that compared the added value of subscores over total scores, we extend this logic to Elo-based skill tracking. Using data from Math Garden, we analyze whether subdomain-specific ratings improve the prediction of students' next responses compared to a unified domain rating. We consider several factors, including the correlations between subdomains, the number of subdomains, student activity levels, and the amount of learning growth. The results aim to inform best practices for structuring skill domains in adaptive learning environments.

Extending the Elo rating system for between-item multidimensionality

H. Vermeiren[1]

[1]KU Leuven, Belgium

The traditional Elo rating system (ERS), widely used to model student learning in adaptive educational systems, assumes unidimensionality and independence between skills. This restricts its capacity to handle realistic scenarios where skills are often intercorrelated. While some extensions to the ERS have been proposed to account for between-item dimensionality, their measurement properties remain largely unexplored. This paper investigates the characteristics of these multidimensional extensions and introduces a novel algorithm that integrates inter-skill relationships in a theoretically grounded way. Unlike models that rely on heuristic correlation terms or hierarchical general abilities, our method derives update formulas directly from a transformed measurement model rooted in multidimensional item response theory. Using a latent rotation technique, we enable the update of correlated skills even when they are not directly practiced. We compare the proposed algorithm to methods from the literature through a simulation study, evaluating prediction accuracy, convergence speed, bias, variance, and robustness to correlation misspecification. Our analysis highlights the trade-offs between model interpretability, theoretical soundness, and practical performance, underscoring the need for continued refinement of multidimensional ERS approaches for better support of personalized learning.

Elo vs. Urnings: A comparison of the measurement properties of different rating-based algorithms in adaptive learning systems

B. Gergely[1, 2]

[1]Tilburg University, Netherlands
[2]Károli Gáspár University of the Reformed Church, Hungary

An Adaptive Learning Systems (ALS) algorithm tracks the performance of students and selects appropriate items based on their activities. Among these algorithms rating systems became appealing, due to their simplicity. The two most prominent rating systems are the Elo rating system (ERS) and the Urnings algorithm (UA). The UA was developed to overcome the theoretical issues of the ERS, by rescaling the latent variables and discretising the updating scheme. In this study we compared the measurement error and convergence speed of the algorithms in simulation, to see whether a discretised nature of UA would result in a slower convergence. First, we assumed that item difficulty estimates are available and compared a simplified ERS and UA, where only the student estimates are updated. The ERS required less item-response to converge, compared to UA. In the second simulation, we monitored the items over time. Here, the ERS could not be used since the estimates are diverging when items are selected adaptively, thus we used the novel Parallel ERS. The UA outperformed the Parallel ERS algorithm, providing faster convergence. In summary, when items difficulties are available the ERS provide accurate ability estimates with less item-responses, whereas if items are updated UA is superior.

Keynote Speech

9:00 - 9:45

Assessment as a Tool for Liberation: Come Dream with Me

J. Randall[1]

[1]University of Michigan, United States

This talk reimagines educational assessment as a transformative force for justice and liberation, rather than a mechanism of oppression and marginalization. Drawing on critical pedagogies of care (Noddings) and discomfort (Boler), I illuminate how assessment systems can elevate learners' critical consciousness, respond to the expressed needs of marginalized communities, and affirm diverse ways of knowing. Through personal narrative (i.e. storytelling) and scholarship, I describe the historical and ongoing harms perpetuated by conventional assessment practices that center whiteness; and advance a liberatory framework for assessment that disrupts these practices. I propose a set of principles, framed within a shared responsibility for justice, that center the voices and lived experiences of rights-holders and actively seek to disrupt structures of inequity. Ultimately, I invite my colleagues to dream with me about assessment systems that not only measure academic achievement, but actively foster growth, social justice, inclusion, and liberation for all learners.

Assessment Cultures VII

9:45 - 10:15

The role of Irelands National Assessments in policy and curriculum development.

J. Kiniry[1, 2], S. Nelis[1]

[1]ERC, Ireland
[2]Dublin City University, Ireland

Irish pupils' strong performance in international studies reflect, in part, the cumulative results of curriculum review and reform guided by robust National Assessment data. This presentation explores the role of Ireland's National Assessments programme in shaping primary education policy and curriculum from its inception in 1972 to the present. The National Assessment programme supplies the Department of Education with regular, curriculum-based measures of reading and mathematics achievement. Initial National Assessments established benchmarks, and subsequent cycles have been informed and inform policy, shaping literacy and numeracy strategies and defining targets. The 2009 National Assessments contributed to the evidence base for the 2011 Literacy and Numeracy strategy, while recent National Assessments highlighted progress and persistent equity gaps, prompting revised policies and initiatives. National Assessment findings have informed school self-evaluation frameworks, digital learning strategies, disadvantaged-school programmes, STEM policy and the development of the recent primary language and mathematics curriculums. National Assessment cycles in Ireland continue to guide policy formulation, ensuring that curriculum priorities, resource allocation and accountability mechanisms remain aligned with student outcomes and educational objectives. In this way the National Assessments in Ireland provide an example of how assessment can both be influenced by, and influence, educational policy.

Informing the Future, Learning from the Past: Reflections from 25 years of empirical research on Leaving Certificate assessment in Ireland

P. Lehane[1], M. O'Leary[1], G. O'Connor[2]

[1]Dublin City University, Ireland
[2]Independent Researcher, Ireland

Despite advancements in how we live and work and the subsequent impact of this on teaching, learning and assessment, the Irish Leaving Certificate (with the exception of what occurred during the acute years of the COVID-19 pandemic) has remained largely unchanged since its inception. As the Leaving Certificate (LC) approaches its centenary, it is timely to reflect on the research conducted to date and consider future directions for policy and study. This research aimed to identify and synthesise the broad range of empirical literature on the LC using Arksey and O'Malley's (2005) six-step scoping review methodology. Following this approach, 107 documents from up to 2024 were included in this study. The review identified six key categories of research: Curriculum Development and Assessment Reform, Maths and Science Education, Fairness, Transition to Higher Education, and Emergency Measures. The findings highlight both areas of sustained research interest as well as notable gaps, particularly around the effects of reform on student learning. By documenting the research landscape, this review aims to support more informed and future-oriented policy and educational reform for the LC.

Empowering Student Agency: A Didactical Model for Designing Learning and Assessment in Technology-Rich Classrooms

Ø. Gilje[1]

[1]University of Oslo, Norway

Despite significant advancements in curriculum and assessment design over the last four decades, few didactical models incorporate how students can demonstrate their knowledge across varied modes using a wide range of semiotic technologies, including artificial intelligence. This paper introduces a didactical model with six choices for curriculum and assessment design in technology-rich classrooms. The choices enhance students' agency through a deeper understanding of the dynamic interplay between learning, assessment, semiotic modes, and digital technologies.

These didactical choices are visually arranged in two triangles, one emphasizing the relationships between assessment criteria, semiotic modes, and technologies. By allowing students to choose how they demonstrate their competencies in varied and authentic ways, the model serves as a toolkit for educators to develop and refine assessment practices, empowering students to make informed choices about their learning.

Drawing from empirical evidence gathered from two extensive research projects in Norway focused on assessment for learning, the proposed model empowers educators to cultivate authentic assessment practices that enhance student agency in technology-rich classrooms. The implications of this model extend to pre-service teacher training, equipping future educators with the tools necessary to navigate the complexities of learning environments enriched with digital technologies and artificial intelligence.

Artificial Intelligence and Assessment VII

9:45 - 10:15

Authoring with AI using a human in the loop approach – is this the future of exam authoring?

K. Evans Johns[1], F. Lawrence[2]

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

Generative Artificial Intelligence has opened up possibilities for the creation of exam material. As an international awarding body seeking to create additional exam content, the potential offered by AI needed to be investigated. A trial was commission to test the capabilities and the output of AI when used with a human-in-the loop approach.
A range of subjects were involved. The output of the first version of the trial was reviewed by external subject matter expert advisers who viewed the exams ‘blind’ – without the knowledge that AI had been involved in the generation of the material.

A second iteration of the trial followed with a more tightly-focused specification and the use of enhanced AI models whilst still integrating the human-in-the-loop approach with the same authoring teams.
The outputs of this trial will be tested by groups of international teachers who will view these exams alongside traditionally authored exams, without knowing that AI was used. The outcome of these focus groups will be shared with the conference and this paper will present further findings on the quality of AI assisted exam content produced using highly trained AI models and highly skilled and experienced groups of human exam authors.

Maintaining test integrity through integration of human judgement and AI systems

S. Hughes[1]

[1]Pearson, United Kingdom

In high stakes assessment, maintaining test integrity is crucial to ensuring that test scores are trustworthy and deserving of public confidence. Challenges to test integrity are nothing new, but they are evolving with the increasing adoption of automated scoring, the proliferation of test preparation advice in social media spaces, and the rapid development of generative AI technology. Further, test takers are facing intense pressures to achieve top scores to access study, work, or immigration opportunities that could change the course of their lives. This pressure can unfortunately motivate some test takers to resort to shortcuts to achieve higher scores or test preparation behaviours that are not conducive to learning. Such behaviours not only threaten the credibility of the assessment but also undermine the validity of the scores awarded. This environment means that testing organisations must constantly innovate to address rapidly evolving behaviours so that test scores carry meaning and can be trusted.

This presentation focuses an innovative solution implemented by a high stakes assessment to address these challenges by strategically integrating AI-based automated scoring systems with human oversight. The balance of AI capabilities and nuanced human judgement is crucial to address these evolving challenges effectively and responsibly.

International Assessments II

9:45 - 10:15

Do students respond more inconsistently to low-stakes conditions? An experimental study manipulating the stakes of the assessment

M. Michaelides[1], E. Konstantinidou[1]

[1]University of Cyprus, Cyprus

Careless responding on surveys undermines the quality of educational assessment data, specifically through inconsistent responses on mixed-worded scales. This study experimentally investigates how motivational conditions (low-stakes vs. high-stakes) affect response inconsistency and response times. In addition, the study explores how personality traits, specifically conscientiousness, and attitudes towards survey participation relate to inconsistent responding, particularly in low-stakes situations. Data collection is currently ongoing with the target sample being 60 participants. So far, 57 university students have completed an online questionnaire individually administered in the lab, including several mixed-worded items, presented under high-stakes and under low-stakes conditions in a counterbalanced within-subjects design. It is hypothesized that inconsistent responding and shorter response times will be more prevalent in low-stakes conditions. Furthermore, higher levels of conscientiousness are expected to predict lower inconsistency under low-stakes conditions. Response inconsistency will be assessed using Mean Absolute Difference scores and Mahalanobis distance. The findings are anticipated to contribute to theories of motivation and survey-taking behavior, providing practical recommendations for improving data quality and assessment design in low-stakes educational contexts. This study is preregistered on the Open Science Framework.

Is two years of teaching for Open Book Exams enough to change teacher practice?

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

[1]International Baccalaureate, Netherlands

In 2021, a literature review on the impact of Open Book Exams (OBEs) revealed that evidence of positive backwash of OBEs on learning, teaching, student performance and wellbeing varied and that even less was relevant to high-school level implementation in an international context. Since 2022, authors have collected multi-wave data from students, teachers and school management from about 280 schools worldwide. Comparing experimental and control group data, after one year the quantitative data showed minimal backwash effects on learning or teaching, while focus groups indicated positive experiences. Initial qualitative data from a small sample point towards significant and possibly enduring changes in teaching practices in the second year, once teachers were confident about expectations. This paper presents the impact of different types of OBE after two years' of embedded preparation from the full sample of schools, comprising a large sample of participating students and of teachers and school staff, including those responding after the May 2025 exam session. Combining survey and focus group data from all three response groups, authors will present backwash effects on learning, teaching and exam preparation practices overall and by OBE type, as well as on epistemological beliefs, growth mindset, wellbeing and student performance.

What you say is what you get? An Experimental Study on the Role of Test Introduction in Test-Takers' Motivation, Anxiety and Performance.

D. Van Looy[1], A. Rogiers[1], C. Frijns[1], M. Vansteenkiste[1], J. van Braak[1]

[1]Ghent University, Belgium

Researchers increasingly recognise the role of test-takers' motivation and anxiety in their test-taking performances. Both result from the interplay between the individual test-takers and the characteristics of the assessment context. How schools and teachers frame tests is of vital importance. In an experimental study involving 302 ninth-grade students in Flanders, the role of test introduction across groups of students according to their background characteristics (i.e., gender, educational track, home language and general achievement level) is investigated. Firstly, the experience of an autonomy-supportive introduction appears beneficial for buffering test anxiety and enhancing students' motivation, both of which support better test performance. Secondly, stressing individual test results, appears to trigger test anxiety. Thirdly, a multivariate analysis of covariance revealed that some groups of students are more vulnerable to the effects of particular test introductions. Particularly, introducing stakes for students increased test anxiety for dual-track students and using a controlling introduction-style exhibited particularly lower motivation for students in the academic track. This presentation outlines the implications of the findings, stressing the role of test introduction to foster students' test-taking motivation, lower their test anxiety and improving their performance.

e-Asssesment V

9:45 - 10:15

Exploring the relationship between students' use of ICT and performance in PISA Mathematics, Reading and Science digital assessments

I. Custodio[1], S. Nastuta[2], L. Liu[1]

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

This research explores the relationship between students' ICT skills and assessment performance, an area of significant interest to policymakers and educators. Previous research has yielded mixed findings regarding the impact of ICT usage on student academic achievement, suggesting that this is a complex relationship that may be influenced by various factors including the purposes and quality of ICT use and students' attitudes, confidence, and competencies. A prior study focusing on the relationship between students' digital skills and performance in Mathematics found that availability and frequency of ICT usage in general tended to negatively impact performance while subject-related ICT use, or enquiry-based learning activities, were positively related with performance in PISA (Programme for International Student Assessment) Mathematics digital items. Our research explores the relationship between ICT factors and student performance in digital assessment items across all three PISA domains: Mathematics, Reading and Science. We make use of the comprehensive PISA 2022 ICT questionnaire to investigate this relationship using 12 composite constructs that represent various aspects of ICT availability, usage and student attitudes. Multilevel regressions were used to assess the impact of ICT on students' performance in the digital Mathematics, Reading and Science items taken by 15-year-old students in England.

Catering to All Learners: Adapting Digital Assessments for Students with Hearing and Visual Impairments

G.A. Nortvedt[1]

[1]University of Oslo, Norway

In recent years, awareness of adapting assessments for diverse learners has significantly increased. With the implementation of universal design, educational authorities and test developers aim to create assessments that meet the needs of all students. This presentation will focus on adaptations made to a national digital mapping test in numeracy for nine-year-olds, specifically addressing how items are tailored for students with hearing and visual impairments, as an addition to universal design principles.
The digital assessment incorporates visual and auditory support to accommodate most children and is available in multiple national languages. However, this support alone does not achieve a truly universal assessment. At the age of nine, many students are still developing their reading skills, making textual support—through either full text or read-aloud features—challenging.
To better assist students with hearing and sight impairments, we have included sign-language videos and concrete materials. This presentation will explore how universal design principles influence test design for young learners and will provide examples of the adaptations made to ensure inclusivity and accessibility for all students.

Enhancing measurement in adaptive learning systems with response time data

M. Bolsinova[1], M. Brinkhuis[2], B. Gergely[1, 3], A. Hofman[4, 5]

[1]Tilburg University, Netherlands
[2]Utrecht University, Netherlands
[3]Károli Gáspár University of the Reformed Church, Hungary
[4]University of Amsterdam, Netherlands
[5]Prowise, Netherlands

Online adaptive learning systems (ALS) aim to make low-cost, tailor-made education available to everyone and to improve both the learning process and the learning outcomes. ALS dynamically adapt the level of learning material based on the individual student's abilities. Therefore, obtaining accurate measures of changing abilities is essential. This is, however, challenging because the amount of response data at a given timepoint is limited, which decreases measurement precision. Incorporating response time (RT) data into ability estimation can potentially improve measurement quality, since these data can provide valuable additional information about ability if modelled appropriately. In this study we compare different models that incorporate RT data for improving measurement quality in ALS with each other and with the baseline Rasch model (i.e., when ability is estimated only based on accuracy) using data from an online learning system for primary school mathematics. Regardless of the scoring rule communicated to the students when collecting the data, the best model in terms of prediction of next response was the one with RT-dependent scores for correct responses (higher scores for faster responses) and a constant large penalty for incorrect responses. We discuss the implications of these results for choosing a measurement model in ALS.

Technical, Vocational and Applied Assessments

9:45 - 10:15

Investigating the relationship between problem-solving elicitation, language demands and difficulty

D. Tonin[1], E. Sweiry[1]

[1]Ofqual, United Kingdom

In England, Functional Skills Qualifications (FSQs) in mathematics are qualifications designed to equip learners with practical skills required for everyday life and employment. One challenge for item writers is developing high-quality problem-solving items accessible to test takers with varying reading abilities. Problem-solving items are typically set in context, and the contexts used may be unfamiliar to some learners. These items usually require more text than context-free items, which can introduce additional language demand. To strengthen our understanding of the properties of good problem-solving items we conducted two studies. In the first study, three mathematics professionals rated 104 items for problem-solving elicitation and identified common features of problem-solving items. In the second, the problem-solving ratings assigned in study 1 were correlated with item facilities (obtained in a high-stakes testing environment) and ratings for language demands and context complexity. Items that were rated as both effective at eliciting problem solving and as having low language demands and context complexity were reviewed by subject experts to identify potential implications for practitioners, such as whether guidance for crafting more accessible problem-solving items could be developed. The findings provide insights for writing problem-solving items that are accessible and inclusive for diverse learners.

Stability of Item Response Theory (IRT) equating based on different sample sizes for functional skills mathematics

Z. Rahman[1], R. Harris[1]

[1]City & Guilds, United Kingdom

A pilot study proved the value of using IRT to equate examinations in a vocational setting. The study involved creating three new versions of a mathematics Functional Skills exam with common items linking test forms. However, in an operational setting, the collection of sufficient test responses for robust IRT analysis was a challenge because of time constraints before releasing results. Although, an evaluation of model fit, including a review of item difficulty, discrimination, and differential item functioning provided reassurance that the quality of items was sufficient for IRT analysis, larger cohort sizes are generally recommended.

Therefore, a follow-up study was undertaken to evaluate the stability of IRT pass marks for different sample sizes and provide more confidence for embedding this innovative approach into business practice. It was found that the pass mark remained somewhat stable even at smaller sample sizes. However, the confidence in the outcome of IRT equating generally increased as the model fit improved and error decreased with increases in sample size. This paper aims to present the findings from this study, highlight the main operational considerations and challenges, and contribute to the limited research in this area in the vocational sector.

Formative Assessment IV

9:45 - 10:15

Sustainable Assessment-for-Learning (AfL): Developing Self-regulated Learning through Evaluative Judgement

S. Caspari-Sadeghi[1], J. Bredberg[1], B. Forster-Heinlein[2], M. Proell[2], K.B.L. Jemaie[1]

[1]Østfold University, Norway
[2]Passau University, Germany

Unlike traditional teacher-led assessment that emphasizes measuring mastery of current content, sustainable assessment prioritizes developing students' competence and responsibility to critically judge their own work—a skill enduring beyond immediate learning. This study investigated the development of evaluative judgement in mathematics and its effects on achievement, Self-regulated Learning (SRL), and critical thinking. Using Design-based Research (DBR), we implemented Assessment-for-learning (AfL) interventions, including self-and-peer assessment, rubrics, exemplars, and feedback, across three undergraduate mathematics courses (N=43) in Norway and Germany over one year. Data were collected from three sources: (a) students' academic performance, (b) engagement with learning content, and (c) perceptions. Analysis revealed statistically significant differences between German and Norwegian approaches towards these interventions. The results support the theoretical claims and prior empirical research that AfL activities positively impact different aspects of academic performance such as engagement, self-efficacy, SRL, and deeper conceptual understanding of mathematics. However, impact is moderated by prior mathematics knowledge as well as attitudes towards perceived usefulness of such interventions which are shaped by the assessment cultures. These findings offer valuable insights for higher education instructors and researchers on designing and implementing structured AfL interventions to cultivate critical judgement capacity.

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

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

Transforming teacher feedback: A checklist for supporting student agency in oral reading

K.M. Gronli[1], B.R. Walgermo[1], P.H. Uppstad[1], E.M. McTigue[1]

[1]University of Stavanger, Norway

Teachers' supportive feedback is essential for young students' reading development. When providing feedback on oral reading, teachers often focus on measurable aspects like decoding, overlooking broader competencies such as comprehension and self-belief, which are also crucial for reading development. Despite extensive research on decoding and fluency in reading assessment, little is known about how formative feedback can intentionally support student agency in early reading.

This study investigates how a 7–9 week feedback intervention designed to support teachers' feedback on decoding and comprehension can also foster student agency in reading. Using a mixed-methods within-subjects design, 52 primary school teachers assessed and suggested feedback on two recorded cases of student readings before and after the intervention, while a subsample of students self-reported on agency. Quantitative changes were analyzed using GLMM with a Poisson link function, and teacher narratives were analyzed thematically.

Findings indicate that decoding remained the central aspect in teachers' assessment, while feedback became more multifaceted, emphasizing motivation, comprehension, and student voice. Student reports and teacher reflections both indicated increased support for agency. The intervention represents a cost-effective and minimal approach to promoting inclusive, student-centered assessment practices. Implications for formative assessment and teacher feedback in early reading instruction are discussed.

Inclusive Assessment II

9:45 - 10:15

Qualifications to Support Curriculum for Wales – The new National 14-16 Qualification Suite

O. Stacey[1]

[1]Qualifications Wales, United Kingdom

In 2022 Curriculum for Wales (CfW) was introduced. CfW represents a shift in the way the curriculum is defined, moving from highly prescribed to purpose led. This approach creates tensions with qualifications with their need for greater levels of prescription to make requirements transparent.

An additional challenge in realising the benefits of CfW is the influence that high-stakes qualification assessment can have on teaching and learning, such as narrowing the curriculum.

At Qualifications Wales (the regulator) we develop the high-level content and assessment requirements for qualifications. To ensure the benefits of CfW are realised and the full breadth of the curriculum assessed, we have designed a comprehensive new qualifications suite (National Qualifications).

Alongside reformed GCSEs we have created new qualification brands including Vocational Certificates of Secondary Education (VCSEs) and Skills for Life and Work qualifications.

The process of designing National Qualifications has involved extensive research and engagement with stakeholders. Throughout, we have balanced competing concerns of assessment reliability and comparability with the need for appropriate flexibility to enable assessments to reflect the different approaches taken to curriculum design.

This presentation outlines the key design features of National Qualifications and how the findings from research with stakeholders has informed their development.

Weak Signals and Wild Cards: An anticipatory framework for the validation of emergent immersive, interactive, adaptive and sensor-based assessments

V. Aryadoust[1], B. Maddox[2, 3]

[1]Nanyang Technical University, Singapore
[2]Assessment Micro-Analytics Ltd, United Kingdom
[3]Digital Education Futures Initiative, Cambridge, United Kingdom

Assessment validation has developed in relation to particular contexts of practice, and technological modes. In the context emergent digital technologies - including sensor-based, process oriented, interactive and adaptive designs, and the application of computational psychometrics and artificial intelligence we have to ask: to what extent to the axioms of assessment theory and organising concepts and principles that used to underpin assessment design and validity practice hold in relation to emergent, and future assessment designs and assessment practice? The paper draws on anticipatory Futures Thinking and French anthropology of technology (Leroi-Gourhan 1943, Simondon 2017; Steigler, 1994), which provides a powerful theoretical and explanatory lens through which we can answer such questions. The phrase – 'weak signals and wild cards', indicates the relevance of anticipatory futures think though, imagine and consider radically different assessment possibilities and scenarios in rapidly changing and highly uncertain technological, ecological and political contexts - and the increased significance of digital 'sensor' technologies in new generations of immersive, interactive and adaptive assessment designs. Leroi-Gourhan's (1943) work suggests, to address those themes we have to recognise the 'tendencies' (i.e., distinctive affordances) of technologies, and the cultural, educational, and institutional contexts in which those are made real as technological 'facts'.

Keynote Speech

11:45 - 12:30

Who Tries and When in the Digital Age: Measuring and Modeling Test-Taking Effort through Process Data in Large-Scale Assessments

M. Ivanova[1]

[1]University of Cyprus, Cyprus

Achievement tests aim for valid estimation of proficiency, but inadequate test-taking effort can introduce construct-irrelevant variance and threaten score validity, particularly in low-stakes contexts. Since 2010s, international programs like the Programme for International Student Assessment (PISA) have tried to measure effort, mainly through self-reports. With the digitalization of large-scale assessments, the large amount of process data now allows for tracing examinees' test-taking behavior, avoiding the biases of self-reports. Process data variables, such as response time, have proven valuable in estimating examinee effort on multiple-choice items, but research on constructed-response items – often linked to lower effort – remains limited. The literature on how thresholds can be determined to distinguish effortless from effortful responses also remains inconclusive. While individual characteristics have been widely studied as predictors of effort, family- and school-level factors have not. Cross-national differences in effort and its relationship to performance are evident, but predictions of effort across countries are yet to be explored in depth. This presentation will discuss indicators, thresholds, and predictors of test-taking effort using process data.