Wednesday, 06 Nov

9:00 - 16:30 Pre - Conference Workshops

Workshop 1
Room: Athena (n=60)

9:00 **Understanding and implementing the moderation of school based assessment for high-stakes examinations**
Damian Murchan, *Stuart Shaw*, *Evgenia Likhovtseva*

Abstract: This workshop provides an opportunity to learn about external moderation of school-based assessment (SBA) used in high-stakes secondary school examinations. Presentations and group work help develop participants' critical understanding of moderation approaches and participants can share their own practices. Many jurisdictions employ SBA within secondary qualifications to increase the validity of inferences about students' learning of aspects of curricula difficult to assess using traditional examinations. This, however, presents challenges with reliability, prompting moderation of SBA. Participants will learn about external moderation of SBA internationally. Topics include: Advantages and challenges with SBA; What is moderation and why is it needed? Techniques available and what factors (e.g. cost) influence choices made? Supporting moderation through planning, professional development, and communication with stakeholders; Intended and unintended impacts of using different approaches. This workshop is based, in part, on a recent review by the presenters of moderation practices in 13 jurisdictions, including interviews with officials. Participants will gain insights into practical issues in conceptualising and implementing moderation in a variety of jurisdictions. Topics will be of interest to assessment professionals and graduate students regardless of location internationally and will provide insights for jurisdictions contemplating introducing externally moderated SBA or wishing to review their current practice.

Workshop 2
Room: Leda (n=60)

9:00 **Optimising the construct validity of test items**
Ezekiel Sweiry

Abstract: The purpose of this workshop is to explore key themes and principles, from both research and practice, relating to how the construct validity of test items can be optimised. Construct validity is taken to refer to the degree to which items assess the underlying theoretical constructs they are intended to measure. Session 1 will consider the key threats to validity posed by different selected and constructed response item formats, and explore the extent to which different levels of thinking can be elicited through these item formats. Session 2 will explore the key features of test items that impact on validity, including language demands, the use of context and diagrams and illustrations. Session 3 will focus on how the mode of assessment (digital or paper) can alter the construct being assessed, and explore the potential for digital assessments to enhance construct validity. The session will also consider how research studies can be set up to further enhance our understanding of the factors affecting item validity. The workshop will include practical activities and authentic example questions throughout to exemplify key points. Opportunity will be provided, across all three sessions, for participants to share and discuss insights and challenges from their own practice.

Workshop 3
Room: Aphrodite A (n=50)

9:00 **An Introduction to the Generalized Kernel Equating Framework with Applications in R**
Jorge Gonzalez, *Marie Wiberg*, *Alina von Davier*

Abstract: The aim of equating is to adjust the score scales on different test forms so that scores can be comparable and used interchangeably. This is extremely important to provide fair assessments to all test takers. The goals of the pre-conference workshop are for attendees to be able to understand the principles of equating, to conduct equating, and to interpret the results of equating in reasonable ways. Emphasis will be given to the new Generalized Kernel Equating (GKE) framework as described in the forthcoming book "Generalized Kernel Equating using R" written by the instructors (Wiberg, González, von Davier, 2024). Different R packages will be used to illustrate how to perform equating when test scores data are collected under different data collection designs. Traditional equating methods, and both kernel equating method and item response theory (IRT) equating methods under the GKE framework will be illustrated. The main part of the training session is devoted to practical exercises in how to prepare and analyze test score data using different data collection designs and different equating methods. Expected audience includes researchers, graduate students, and practitioners. An introductory statistical background as well as experience in R is recommended but not required.

Workshop 4
Room: Aphrodite B (n=50)

9:00 **Breaking barriers for all test-takers**
Caroline Jongkamp, *Helen Claydon*, *Thomais Rousoulioti*, *Renika-Irini Papakammenou*

Abstract: It is often the case that diversity and inclusion are afterthoughts when an organisation is evolving its summative e-assessment offering. This workshop will provide an engaging opportunity for collaboration with peers, to consider the perspectives of a range of test-takers. Thought-provoking discussions will equip participants with areas to take away and integrate in their future work practices. The premise for the workshop is that participants set new priorities to develop e-assessments and assessment services to support test-takers with a range of different forms of special educational needs and disabilities (SEND) and culturally diverse backgrounds. The workshop will focus on the test taker and consider how all parties in the test process (test developers, test administrators, teachers, school administrators) can support fair testing practices. The participants will work in groups as test-takers with different needs and explore how e-assessment can break barriers for all test-takers. This workshop is led by members of the AEA-Europe eAssessment and Inclusive Assessment SIGs. No prior experience of e-assessment or inclusive assessment is needed.

Workshop 5
Room: Christian Barnard (n=200)

9:00 **Assess your assessment**
Bas Hemker, *Cor Sluijter*

Abstract: This workshop will provide participants with all the tools needed to formally assess educational assessments, either computer-based, paper-based, or through an assessment system of their choice. Assessing the quality of your own assessment serves as an instrument for quality assurance of the assessment. It also helps to communicate the quality of assessment, to ensure accountability to end users (students, teachers, schools, policy makers) and to the general public. During the workshop participants will actually evaluate the instrument of their choice via a validated review system. This provides them with valuable information about the quality of their educational assessment or assessment system and can help them to further improve its quality. Step by step, the workshop will provide participants with guidelines on how to proceed with the development of new instruments, including the use of psychometrics, thereby increasing the chances of efficient production of high-quality instruments in the future. The workshop is intended for participants who have some experience with the development of assessment, who want to know more about ways to evaluate (and improve) their assesments. The required previous knowledge is the knowledge used and obtained with test development.

Workshop 6
Room: Hermes (n=24)

9:00 **Evaluating Impact in the Context of Educational Assessment**
Brigita Seguis, *Hanan Khalifa*

Abstract: The aim of the workshop is to provide a comprehensive overview of key concepts, methodologies, and best practices for assessing the impact of assessments, educational programmes, interventions and policies. Designed for professionals involved in educational assessment, research and policy, the workshop will equip participants with the knowledge and skills needed to conduct impact evaluations and provide evidence-based decisions to improve educational outcomes. The workshop will begin with an introduction to the importance of impact evaluation in educational assessment, focusing on key definitions (e.g. washback, impact, consequential validity), concepts (e.g. input, output, outcome) and evaluation frameworks and models (LogFrame, Theory of Change, Kirkpatrick's model). Participants will then explore different types of impact methodologies, covering experimental, quasi-experimental and non-experimental approaches, based on practical examples from various educational settings. The essential steps involved in designing and conducting impact evaluations will be covered in the second part of the workshop. Participants will learn practical strategies for formulating evaluation questions and hypotheses, defining evaluation indicators, selecting data collection instruments and engaging stakeholders throughout the process. At the end of the workshop, participants will leave with a deeper understanding of impact evaluation principles and practices that they will be able to apply in their own settings.

Workshop 7
Room: Zeus (n=18)

9:00 **Introduction to multilevel modelling using large-scale assessment data**
Anastasios Karakolidis, *Vasiliki Pitsia*

Abstract: International and national large-scale assessments, such as TIMSS, PIRLS, PISA, and NAEP, play an important role in informing educational policies and practices across countries. Such assessments provide rich but complex data due to their assessment and sampling designs. It is important to be aware of these complexities in order to analyse large-scale assessment data correctly and interpret results appropriately to inform policy and practice. This workshop offers an accessible theoretical and practical introduction to multilevel modelling, a technique that allows for the appropriate analysis of large-scale assessment data and offers significant advantages compared to other single-level techniques (e.g., examination of interactions between student- and school-level factors). Specifically, the workshop presents key concepts and design features of large-scale assessments that are relevant to multilevel modelling (e.g., cluster sampling, weights), introduces attendees to the theory behind multilevel models, considers issues from a practical perspective to support data preparation and the selection of modelling techniques, and engages attendees in the application of multilevel modelling and the interpretation of its results. Upon completion of the workshop, attendees are expected to have a good understanding of key aspects of large-scale assessments and multilevel modelling and be able to run their own multilevel models.

19:00 - 20:30 Welcome Reception for all participants

Location: Coral Beach Hotel & Resort

Thursday, 07 Nov

8:00 - 9:00 Registration

8:00 - 9:00 Meeting of SIG SC Chairs
Room: Hermes (n=30)

9:00 - 9:45 Welcome Addresses
Room: Akamas A&B (n=550)

Prof. Therese Hopfenbeck (President AEA-Europe)

Dr. Athena Michailidou, Ministry of Education, Sport and Youth

Prof. Elena Papanastasiou, University of Nicosia

9:45 - 10:30 Keynote Speech
Chair: Therese Hopfenbeck
Room: Akamas A&B (n=550)

Dr. Yiasemina Karagiorgi, Head of Educational Research and Evaluation (CERE) of the Cyprus Ministry of Education and Culture

10:30 - 11:00 Coffee Break

Foyer outside Akamas Room

Opportunity to visit SIG Banners

11:00 - 12:30 Open Paper Session I

Artificial Intelligence I
Chair: Bas Hemker
Room: Akamas A&B (n=550)

11:00 **Exploring the Potential of Artificial Intelligence on Educational Assessment: Insights from the Student and Educator Perspective**
Agni Stylianou Georgiou, Elena Papanastasiou

Abstract: This study aims to explore the perspectives of secondary school students, their teachers, and academics on the potential of AI integration in assessments in schools. Through a mixed-methods approach, involving online questionnaires and thematic analyses, the findings reveal students' envisioning AI's role in assessments as one that enhances automaticity, provides educational support through personalized adaptation, and makes the assessment process more enjoyable through gamification. Interestingly, students also perceive AI as a means to achieve greater objectivity in grading, contrasting with a minority who value the subjective judgment of teachers. Teachers and academics echo similar themes, recognizing both the opportunities and challenges that AI presents, including concerns about the diminishing human touch in education. This study highlights the importance of incorporating diverse perspectives in the discourse on AI in educational assessment, which is essential for harnessing AI's full potential in assessments. The findings contribute to the ongoing dialogue on AI's role in education, offering valuable insights for future applications and research in AI-enhanced educational assessment.

11:30 **A Norwegian case study of student teachers' perceptions and experiences of AI and AI-assisted feedback – mapping diverse users**
Siv Gamlem, Joshua McGrane, Synnøve Moltudal, Sundance Zhihong Sun, Christian Brandmo, Therese N Hopfenbeck

Abstract: This study investigates student teachers' familiarity, perceptions, and attitudes regarding Artificial Intelligence (AI), how AI and AI assisted feedback (AIF) empower their teacher agency, and what AI-related challenges they perceive as future educators in the classroom. Student teachers enrolled at a 5-year teacher education program for primary and lower secondary school at a University College in Norway were invited for participation (N = 216). A mixed-methods design was conducted. Data were first collected through a survey (n=209, 97 % response rate) and followed up by individual interviews (n=11). Results reveal that a substantial majority of participants reported having used an AI application and knowing about AI, although none considered themselves experts. In terms of perceptions and attitudes, participants saw pragmatic advantages of AIF in lesson planning and finding content but were equally cautious about its trustworthiness and the broader implications for human-centred education. Based on the analysis of the interview data regarding how student teachers perceive AI and AIF to support learning, feedback, assessment, and their teaching, we revealed diverse viewpoints influenced by their personal use of the technology. The comparison between experienced and novice users revealed a divide in expectations for AI's and thus AIF's role in education.

12:00 **The Use of Artificial Intelligence in Qualifications: Perspectives on Regulation**
Vasile Rotaru

Abstract: Artificial Intelligence (AI) is gaining popularity across various sectors and its integration into systems and decision-making processes has sparked considerable interest and investment. However, alongside its transformative capabilities, concerns surrounding trust and accountability have emerged as a significant theme. With AI increasingly employed for various tasks, questions about the reliability of AI-driven decisions and their ethical implications have also become prominent. In response, there is a growing consensus that proper regulation of AI is necessary. Clear guidelines can help reduce risks, ensure AI's benefits are realised and foster trust among users and the public. This presentation will draw upon various ideas expressed in the literature on the responsible and acceptable use of AI and explore key considerations in regulating AI use in qualifications. It will examine some of the challenges and opportunities inherent to regulating AI in qualifications. It aims to initiate a discussion rather than offer definitive answers. To enrich the discussion, the presentation will incorporate findings from a qualitative study conducted in Wales. The study interviewed stakeholders involved in qualifications, such as teachers, policymakers, and awarding bodies. Their insights will help understand how regulation should affect the use of AI in qualifications.

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

Chair: Alex Scharaschkin
Room: Aphrodite A (n50)

11:00 **The use of machine learning in predicting students' exam grades: a case study and discussion of ethical implications**
Georgie Billings

Abstract: This paper discusses an experiment with using the random forest machine learning as a mechanism for predicting student grades. Random forest is a commonly used machine learning algorithm trademarked by Leo Breiman and Adele Cutler . It solves classification problems (in this case, which grade does each student 'belong' to) by using many randomly generated individual decision trees as an ensemble. Each tree in a random forest will make a class prediction, and these are then combined, with the most 'popular' class in the forest becoming the model's prediction. This avoids the issues around overfitting that hamper individual decision trees. The research focuses on Grade 12 Mathematics examinations for a subset of students in Kazakhstan. It involved 'feeding' the algorithm feature variables, such as candidates' average performance in internal summative assessments and scores in earlier exams such as end of year tests or IELTS exams. The model was then trained on the actual grades of students in their Grade 12 Mathematics exams and was able to realise an 83% accuracy in predicting grades This session is intended to present the research findings of the case study, and to generate and encourage discussion of the implications of using such methods.

11:30 **Implementing Innovative Assessment Methods in the Teaching Practicum: Exploring the Insights of Examiners in Malta**
Josephine Deguara, Josephine Milton

Abstract: This study explores the insights and perspectives of teaching practicum examiners in a Maltese initial teacher training institution about their assessment experiences during the COVID-19 pandemic. With traditional in-person observation assessments disrupted, examiners had to adapt to remote evaluation methods, such as assessing video recordings of teaching practicum lessons or activities. Modes of giving and sharing feedback about the assessments were also adapted through the use of video conferencing, to ensure pandemic mitigation measures were upheld. The study employs a qualitative interpretive research paradigm, collecting data through textual narratives from some teaching practicum examiners. Findings reveal that the transition to remote evaluation presented significant challenges, including increased time demands and limitations in gaining comprehensive insights into teaching practices. Examiners emphasised the irreplaceable experience of face-to-face assessments, highlighting the nuanced observations and interactions that in-person evaluations provide. However, the study also identifies the potential for a hybrid assessment approach that combines remote and face-to-face evaluations. This model offers benefits such as accessibility, flexibility, and efficacy, accommodating diverse learning styles and contexts. The results suggest that while remote assessments cannot replace face-to-face ones, a hybrid approach can enhance the overall effectiveness of teacher education programmes.

International Assessments I
Chair: Dario Pirotta
Room: Leda (n=60)

11:00 **Improvement of contextual questionnaire scaling using machine learning to identify unusual responses**
Tim Friedman, Dulce Lay

Abstract: International large-scale assessments such as ICCS 2022, gather substantial amounts of student background data which are essential for informing educational policies and practices. However, the validity and reliability of scales produced using such data can be compromised by the effort of respondents, impacting the psychometric properties of latent constructs being measured. Our study applies machine learning techniques to analyze ICCS data across three cycles to re-scale and enhance the psychometric quality of student background questionnaire scales. We identify students whose responses deviate from typical patterns, indicating careless, inconsistent, or insufficient effort. Anomalies are detected and characterized through supervised and unsupervised machine learning algorithms to investigate similarities and differences in response styles across cycles. Analyses focus on questions related to students' expected electoral participation, expected political participation, legal and illegal protest activities, citizenship self-efficacy, trust in civic institutions, and endorsement of gender equality. Results show a significant improvement in the psychometric properties of six out of eight scales, and the identification of sub-groups of the population who are more likely to provide illegitimate responses. We believe the results of our study are important in maximizing the use of legitimate contextual data collected from international large-scale assessments.

11:30 **Are student and parent reading attitudes and behaviours related? Evidence from PIRLS 2021 for Ireland**
Vasiliki Pitsia, Sarah McAteer, Emer Delaney

Abstract: Although parents' role in shaping children's academic outcomes has long been acknowledged, research examining their role in shaping children's attitudes and behaviours is scarcer. Given parents' central role in their children's development and the importance of children's attitudes and behaviours in academic and other terms, this study examines the relationships between students' and parents' reading attitudes and behaviours and the extent to which these relationships vary by student gender and socioeconomic status. Data from the Progress in International Reading Literacy Study (PIRLS) 2021 on 4,663 fourth-grade students in 148 schools in Ireland were analysed. Results indicated that students' liking of reading, the extent to which they feel confident in reading, and the time spent on out-of-school reading were statistically significantly and positively related to parents' reading attitudes and behaviours. Some of these relationships varied by student gender and all were stronger among students from higher socioeconomic backgrounds. Latent class analysis based on the reading attitudes and behaviours of students and their parents, respectively, provided further insights. Implications for policy and practice are discussed, along with the benefits of making more extensive use of the non-cognitive data collected through international large-scale assessments.

12:00 **Do students respond inconsistently on mixed-worded scales in the PISA 2022 questionnaire? Evidence across six educational systems**
Evi Konstantinidou, Militsa Ivanova, *Michalis Michaelides*

Abstract: The exploration of the respondents' behavior in survey assessments which include mixed-worded scales has gained considerable attention recently. We investigated the prevalence of inconsistent responders in two mixed-worded scales: the 6-item Sense of Belonging and the 10-item Stress Resistance Scales, across six countries (i.e., Australia, Brazil, Denmark, Greece, Saudi Arabia, and Singapore) participating in the Programme for International Student Assessment (PISA) 2022. The results revealed that inconsistent responders were 4.3% in the first and 16% in the second scale with sizable cross-country differences. Logistic regression analysis revealed that reading achievement was positively and consistently linked to consistent responding across countries and scales. Other characteristics such as gender, effort, and time on screen were found to be important predictors in some of the country samples. The study discusses implications for the use and development of surveys with mixed-worded scales emphasizing the need to consider contextual factors like the position of a scale within a survey and cognitive and demographic characteristics of survey participants across different cultural settings.

Higher Education & Assessment I
Chair: Rebecca Conway
Room: Akamas C (n=200)

11:00 **Assessing Students' Legal Literacy in Higher Education Using Computer-based assessment**
Ksenia Tarasova, Daniil Talov, Sergei Tarasov

Abstract: The modern employment market puts forward the requirements for university graduates to have not only professional knowledge, but also to form universal competencies and various types of "new literacy" relevant to the modern world. A special place among them is taken by legal literacy. We utilized Evidence-Centered Design to develop a computer-based assessment instrument to assess Legal Literacy (LL) of 1st year undergraduate students of non-legal specialties. By performing 9 scenario-based tasks the respondent has the opportunity to demonstrate his or her skills, knowledge and attitudes in different areas of law. According to the results of psychometric analysis, it was shown that the instrument for measuring students' legal literacy has good psychometric characteristics, including high reliability and agreement of the tasks with the model, and can be used for scientific research, as a basis for educational interventions and for providing feedback to respondents. As the development of civil society is an essential task, it is important to be able to measure the knowledge, skills and attitudes in this area. The presented approach may be of interest to researchers in the field of legal literacy and developers of assessment tools in an interactive digital environment.

11:30 **In Search of Assessor Identify during the Teaching Practicum: Insights from 'Conversations' between Teacher Educators and Student Teachers**
Michael A. Buhagiar, Deborah A. Chetcuti

Abstract: This presentation focuses on the assessment of the teaching practicum (TP). We are two teacher educators at the Faculty of Education, University of Malta. As part of our work, we lecture mathematics and science student teachers on assessment, prepare them for TP and assess their teaching during TP. Years back, we were instrumental in shifting our Faculty's TP assessment discourse and practices towards formative practices. But now, Faculty is moving back towards summative practices by reintroducing marking/grading. This development necessitates that we redefine our assessor identify to operate within a TP assessment system that is not aligned to our beliefs and practices. Towards this, we decided to hear our students' views about the change in TP assessment and how this would impact their TP-related work and learning. We explored this through small-groups online recorded 'conversations' with them. We are currently analysing the data transcripts using an inductive thematic approach. The indications thus far are that student teachers might have views on TP assessment that do not necessarily reflect those of their lecturers, even when these lecturers are responsible for their assessment literacy, and it seems that this could be attributed primarily to different sets of priorities and values.

12:00 **Unraveling the Self-Feedback Process: Exploring the Black Box of Self-Assessment Through Multiple Studies**
Ernesto Panadero, Javier Fernandez Ruiz, Leire Pinedo, Iván Sánchez

Abstract: This presentation synthesizes findings from two data collections published in four papers to explore the dynamics of self-assessment among secondary and higher education students under various feedback conditions. Involving over 500 participants across different educational levels, the research investigates the influence of rubric-based feedback, instructor feedback, and their combination on self-assessment practices. The experiments utilized think-aloud protocols, direct observation, and self-reported data to elucidate the self-assessment actions—namely reading, recalling, comparing, rating, assessing, and redoing tasks. Findings reveal that while the type and timing of feedback generally enhance the sophistication of self-assessment criteria, the impact varies by educational level and the complexity of feedback. Secondary students demonstrated similar self-assessment profiles to their higher education counterparts, with external feedback sometimes impairing the richer, self-driven assessment processes. Notably, the integration of feedback types was found to be crucial, with combined feedback often yielding the most substantial improvements in self-assessment accuracy and depth. This body of work contributes to the ongoing discourse on optimizing self-assessment in educational settings, proposing a model of self-feedback (SeFeMo) that encourages educational stakeholders to refine feedback strategies to foster deeper learning and self-evaluative skills.

E-Assessment I
Chair: Andrew Boyle
Room: Athena (n=60)

11:00 **Using assessment and response times data to evaluate a digital mock exams service**
Carmen Vidal Rodeiro, Tim Gill, Sarah Hughes

Abstract: Cambridge University Press & Assessment provides a Digital Mocks Service for GCSE and A-Level qualifications. Schools in England and around the world can sign up for the service and deliver digital mocks to their students. The mock exams, available in a range of subjects, are usually based on exams delivered on paper in previous live sessions. The capture and use of data from digital assessments (process data) has the potential to help quality assurance of tests/items, help understand test-takers' behaviours, engagement and motivation, and improve the quality and reliability of assessments. This work made use of assessment and response times data to evaluate the Digital Mocks Service. Its outcomes showed that, for most of the qualifications available in service, reliability was in line with values for reliability in large-scale assessments. Furthermore, item difficulty levels ranged from difficult to fairly easy in a similar way they do in live paper-based assessments and item discrimination indices were acceptable. Analysis of response times data showed that there was no evidence of candidates running out of time and that time spent in the items did not decline as the students progressed through the assessment (i.e., there was no evidence of speededness or students' disengagement).

11:30 **Going digital? The impact of shifting the mode of high-stakes assessments in England on students**
Yasmine El Masri, *Jo Handford*, Harvey Dodds

Abstract: With greater adoption of digital assessments globally, the regulator of qualifications and examinations in England is examining the opportunities and risks of moving high-stakes assessments from pen and paper to the computer screen. The literature outlines many benefits to on-screen assessments in terms performance, levels of engagement and accessibility; however, some of these claims are not backed by robust research. The evidence available is at best mixed and has been primarily carried out in higher education and low-stakes contexts making conclusions less relevant. In absence of a solid evidence base, research was commissioned to build a better understanding of the impact of greater adoption of on-screen assessment on students in England in high-stakes qualifications. The research examined stakeholders' perspectives on the nature and extent of the impact on different groups of students. Interviews were carried out with students, parents, teachers and subject matter experts selected from diverse backgrounds. Findings suggest that, while stakeholders recognised the opportunities that on-screen assessments can provide, they also expressed various concerns around the perceived impact such a change might have on students' performance and experience of assessments as well as the sector's readiness for deploying on-screen high-stakes assessments on a large scale.

12:00 **On-screen high-stakes assessments: Lessons learned from other jurisdictions**
Yasmine El Masri, *Jo Handford*, Harvey Dodds

Abstract: The regulator of qualifications and examinations in England is examining the opportunities and risks associated with moving its pen-and-paper high-stakes assessments to the computer screen through engagement with countries who have progressed further on the journey. A systematic multi-step review of 50 countries led to the selection of eight jurisdictions for in-depth study. Workshops were held remotely with key individuals from relevant institutions within each country, including government organisations and examinations boards. The workshops probed for motivations for moving national assessments on screen; barriers, risks and challenges in deployment; the approach to deployment, including details of implementation plans; and impacts and benefits on various stakeholders. The study suggests that most countries planned to move high-stakes assessments on screen as part of a wider educational reform. Key drivers included equipping students with digital skills as well as potential improvement in efficiency, security and resilience within the system. Challenges to deployment included securing funding, the adequacy of existing school infrastructure, fairness of treatment of students as well as overcoming the resistance of influential stakeholders to change. Countries adopted different approaches to deployment. These will be examined in terms of their relevance to the English high-stakes qualifications context.

Assessment Cultures I
Chair: Emma Walland
Room: Zeus (n=30)

11:00 **Policy and practice in relation to external moderation of School-based Assessment in 13 education systems internationally**
Damian Murchan, Stuart Shaw, Evgenia Likhovtseva

Abstract: This study investigates why and how different jurisdictions conceptualise and operationalise external moderation of School Based Assessment (SBA). A range of concerns about high-stakes examinations at upper secondary level have prompted some systems to incorporate SBA into their system of qualifications. While potentially addressing issues of validity and student stress, SBA raises reliability concerns that can also compromise trust in the qualifications. External moderation is frequently used to allay such concerns. The aim of the study is to identify illustrations of moderation in selected secondary school exit examinations and understand the local contexts that have contributed to the development of such systems. A two-phase sequential survey design was used to explore the variables of interest across 13 jurisdictions. This involved a review of publicly available documentation and interviews with examination officials in 9 examination organisations. Findings reveal how education systems use moderation to ensure consistency of standards within and across schools. A range of approaches were identified, consistent with the literature, but showing a marked mixing of models. Identifying a preferred approach is elusive; rather, decisions are governed by local national contexts, including capacity within the systems. The findings provide useful advice for jurisdictions contemplating introducing externally moderated SBA.

11:30 **Teacher Critical Consciousness in Educational Assessment: Why is it important and how can we develop it?**
Catarina Correia

Abstract: Educational Assessment has a tremendous influence on teaching and learning. In England, teachers and students experience two different and potentially contradictory discourses around assessment. One where assessment is used to position students as capable learners with potential to develop and grow. Another where assessment results are used for the construction of ‘success and failure’. In addition, ideas of meritocracy permeate these discourses overlooking the differential impact that oppressive social structures have achievement. The over emphasis on test-preparation/test-results emphasises the construction of the ‘success and failure’. Teachers and students are not educated to navigate these contradictions, and this has significant negative consequences on their well-being. A review of current conceptions of teacher assessment literacy reveals that whilst some dimensions encourage teachers’ autonomy in their classroom assessment practices, they do not equip teachers to challenge and take political action to address broader structural issues that shape what goes on in the classroom. Drawing on Paulo Freire’s (2021) ideas of critical consciousness as a vehicle for building awareness of and taking action to change oppressive systems, I discuss how teacher assessment literacy can be revisited to include elements of critical social analysis and social mobilisation.

12:00 **Understanding progression and assessment in the context of the new Curriculum for Wales**
Estelia Borquez - Sanchez, Kara Makara - Fuller, *Lesley Wiseman - Orr*, Fiona Patrick

Abstract: The Welsh Government is implementing a new curriculum supporting schools and practitioners in designing their own curriculum and co-constructing assessment approaches aligned with this framework. CfW requires that schools develop a shared understanding of 'progression', i.e., 'developing and improving skills and knowledge over time', and it asserts that the primary purpose of assessment is to support learners in that progression. The Camau i'r Dyfodol project has been funded by the Welsh Government to support education professionals in Wales. Findings indicated some confusion in the system regarding understanding progression and assessment in the new curriculum. As part of our work from phase 1 report (2023), we undertook a narrative literature review to synthesise what is known about the relationship between curriculum, assessment, pedagogy and learning progression. The review found a lack of research on generic learning progression and connections between curriculum, assessment, and pedagogy. As a more specific concept than described within CfW, LPs are models based on research and classroom evidence, describing pathways to help teachers effectively teach and assess. Given that the understanding of progression in CfW differs from the literature, we conclude by discussing definitions across contexts and outlining implications for understanding progression and assessment in the CfW.

Formative Assessment I
Chair: Doreen Said Pace
Room: Christian Barnard (n=200)

11:00 **Dialogues on learning and assessment (DOLA): Attitudes and tensions in assessment practices – in the way for learning and motivation?**
Kathinka Blichfeldt, Kaja Haaland, Ingrid Jacobsen

Abstract: This paper examines the understanding of continuous assessment among newly qualified Norwegian teachers, questioning whether confusion between the purpose of continuous and final assessments may reinforce outdated practices contrary to the aims of recent legislation. These concerns arise from issues in long-term, decentralized competence development partnerships between Inland Norway University of Applied Sciences and educational institutions in Norway. 244 newly qualified teachers took part in half-day seminars aimed at enhancing their understanding of the assessment framework. Through a mixture of workshops and mentimeter surveys, the study explores the experiences and attitudes of teachers within a Norwegian municipality. The seminars focused on assessment culture, exploring the influences on new teacher's assessment practises and their developmental needs post-reform. Preliminary findings suggest that current assessment associations are predominantly negative, emphasizing ranking rather than being a motivational learning tool as intended. This highlights the necessity to develop assessment practices following the 2020 reform to enhance professional learning communities and facilitate quality development in school-based assessment. There's an identified need for a deeper exploration of assessment concepts and practices to align them more closely with learning-focused dialogues, consistent with the Education Act's objectives.

11:30 **What supports high-quality approaches to assessment? Predicting student teachers' competence in educational assessment by personality, motivation, and attitudes**
Christoph Schneider, Christopher DeLuca, Lothar Müller, Andrew Coombs

Abstract: Assessment competence is of paramount importance for teacher education (TE) programmes. Assessment learning is a complex, social endeavour, involving theoretical input, practicums, and reflection, as well as sufficient time. Our paper investigates whether personality measures, motivation, and attitudes predict student teachers' endorsement of 'contemporary' (i.e. non-standardized, formative) approaches in educational assessment. In a longitudinal survey (n=453 student teachers), approaches to assessment, personality, motivation, and attitudes were surveyed at three timepoints. Latent change modelling demonstrates student teachers' progression in 'contemporary' approaches across time in TE. Moreover, multiple prediction shows that the endorsement thereof is higher amongst those having entered TE out of intrinsic motivation and holding positive attitudes towards inclusive classrooms. Data support that student teachers' learning about assessment occurs at slow pace and that motivational features and attitudes facilitate learning about assessment. This study adds further evidence that a new understanding in the assessment preparation of future teachers is needed.

12:00 **Connections between teachers' and students' understandings of continuous and final assessment**
Egil Weider Hartberg, Kari Kolbjørnsen Bjerke, Kjell Evensen, Trude Slemmen Wille, Terje Engh Wiig

Abstract: The Norwegian enactment of assessment, revised in 2020, clearly states that continuous assessment should contribute to the students' 'desire to learn' and that formative assessment should be an integrated part of the students' learning process. The Inland Norway University has maintained a professional partnership with several secondary schools in the Oslo municipality for over four years, focusing on teachers' assessment practices and their underlying understanding for formative assessment. During the last year we have extended the focus to include how a clear understanding of the final assessment process, among teachers as well as students, provide an important frame for the learning process. Our study seeks to understand the relationship between teachers' practices and understanding of continuous and final assessment, and students' understanding and experiences of these processes. As part of the project, we study how teachers and students understand the different role AI represent in continuous/integrated and final assessment. Research questions: 1. How does the final assessment affect students and teachers understanding of formative assessment during the course? 2. How do the students understand and experience their teacher's didactical practice in formative assessment?

Assesment of Practical Skills I
Chair: Sebastiaan de Klerk
Room: Aphrodite B (n=50)

11:00 **How much data to feed to a neural network for autoscoring?**
Anastasiia Beliaeva, Elen Abdurakhmanova, Daniil Talov

Abstract: Complex constructs like reading literacy can not be measured solely by multiple choice tests. To assess such skills, open-ended questions should be included in the testing. However, assessing such format of answers implies increased human resources: the time to check the works goes up, and so do the costs. Another negative aspect of manual assessment is the need to teach raters how to use the criteria to assess the test, which is also relatively costly and time-consuming. There is also an urgent need to check inter-rater reliability in order to make sure the marks for the tests are valid. One possible solution to these apparent drawbacks of manual assessment lies in the development of an automated system. Recently, many Large Language Models (LLMs), which is a type of neural network, have emerged to deal with the task of automated text analysis. Implementation of such pretrained model can, on the one hand, dramatically decrease the time for checking the work, and on the other, reduce the amount of money to be spent on raters.

11:30 **Cross-institutional Clinical Skills Assessment Quality Assurance in Europe, a mutual assessment strategy; are we equipped for it?**
Thomas Kropmans, Eirik Søfteland, Magnus Hultin, Rosemary Geoghegan, Angela Marie Kubacki

Abstract: Introduction Objective Structured Clinical Examinations (OSCE) is a well researched, laborious and costly paper based method of exam delivery, restricting international comparison. Cross-institutional comparison of OSCE Quality Assurance in Europe has never been done and due to widespread electronic assessment analysis is now available. Method: Eight European educational institutions using an electronic OSCE Management Information System confirmed to join a mutual comparison of Quality Assurance outcome. Outcomes were compared including the Standard Error of Measurement (SEM) as well as cut-scores, Pass/Fail score and Global Rating Scores, Cronbach's Alpha and related SEM (68% and 95% CI). Results: The Classical psychometric based SEM varies from 2.8% to 11.2% respectively, whereas the 95% CI equivalent varies from 9.2% up to 22% (on a 0 - 100% scale). The relative SEM from G-theory analysis varies from 3.15% to 7.0% for criterion-referenced marks, and the absolute SEM for norm-referenced marks varies from 3.8% to 7.8% respectively. The 95% CI around the relative and absolute SEMs values varies from 7.3 to 15.3%. Discussion/conclusion: To protect society and to improve educational decision-making the Standard Error of Measurement and associated confidence intervals needs to be embedded in EU assessment strategies to rule out 'false positive Pass decisions'.

Comparative Judgement I
Chair: Tom Bramley
Room: Hermes (n=30)

11:00 **A new Comparative Judgement (CJ) approach: Exploring the potential of criteria-based CJ**
Nicky Rushton, Victoria Crisp

Abstract: Previous comparative judgement (CJ) studies show that some judges find making judgements difficult and lack confidence in their decisions. The lack of transparency regarding the criteria that judges use may also be a concern. Some studies have provided reference documents such as importance statements to support judgements; however, judges still differed in the criteria used. The current research explored whether asking judges to make separate comparative judgements for several broad criteria better supports judgements. Ten assessors completed a holistic exercise where they judged which exam script in each of a series of pairs showed better overall performance, and a criteria-based exercise where they judged which script in each pair was better with regard to each Assessment Objective (AO). The order of the exercises was counterbalanced. Assessors completed a workload questionnaire and an experience questionnaire after each exercise, and a final experience questionnaire. Results suggest that criteria-based CJ is a plausible alternative CJ approach. Script ranks correlated strongly with those from holistic CJ and the approach provides greater reassurance that key constructs inform judgements, with some evidence of more script features being considered. However, criteria-based CJ did not strongly improve perceived ease of making judgements or assessors' confidence.

11:30 **Evaluation of markers' performance involved in the process of marking written works on open ended items in high-stakes examinations.**
Ali Mahmudov, Sarkhan Guliyev, Fuad Ahmadov, Elmir Shirinov

Abstract: The State Examination Center (SEC), with more than 30 years of experience in the assessment field, has been successfully conducting school-leaving examinations and the admission process to higher education institutions. In connection with the application of the new-curricula 2008, SEC developed a new technology to assess the knowledge and competencies of students prepared on the basis of the new-curricula. To assess the knowledge and skills of students studying under the new-curricula, SEC has developed a new format of school-leaving examinations starting from 2017, as well as a new format of university entrance exams, where open-ended items are used in different subjects. To mark students' written works on open-ended items, the SEC has developed and uses a new assessment technology called "E-Marking" system, which allows marking online (or on-site). Written works are assessed by trained markers. Markers undergo a calibration process before marking written works on each item. In addition to the calibration process, an analysis (audit) of the markers' performance during and after the marking process is carried out based on several indicators. The paper will also present the findings and implications of the analysis of the inter-rater reliability of the marking process based on Krippendorff's alpha and generalizability theory.

12:00 **The future of assessment with natural data capture**
Kemran Mestan

Abstract: Increasing quantities of data associated with teaching and learning can be collected naturally – i.e., Natural Data Capture (NDC). NDC can be achieved through online learning platforms, diminishing the need to administer discrete assessments. A small proportion of online learning could yield a broader snapshot of student abilities, than interspersed testing. NDC reduces the cost, time and distortion associated with traditional testing, but challenges remain. Some online learning exercises may lack a grounding in rigorous curriculum and assessment standards. Further, a cacophony of online resources may cause incoherence. Regardless how much data is collected, two related questions remain: 1. What online learning activities are appropriate for a given students' abilities? 2. What does the data from online learning activities reveal about students' abilities? With the aim to support education systems become more coherent the Australian Council for Educational Research (ACER) has developed Learning Progressions - continuums that maps key stages of development, underpinned by numerical scales. The Pairwise Comparison Method (PCM) calibrates items on such scales. This method could be applied to online learning data. Treating online learning exercises like assessment items. Thereby, mapping student ability to undertake online learning activities to learning progress.

12:30 - 13:30 Lunch

Armonia Restaurant

13:30 - 15:00 Discussion Groups

Discussion Group 1
Room: Akamas C (n=200)

13:30 **Is artificial intelligence considered a helpful ally or a potential antagonist in the field of assessment research?**
Dan-Anders Normann, Julie Leonardsen, Gabriel Cipriano, Estelia Borquez Sanchez, Shakeh Manassian

Abstract: The Post Graduate Student and Early Researcher Network aims to create a supportive environment for all its members. As part of this goal, we invite current members and interested conference participants to join us for a discussion session to explore the theme Advances in Educational Assessment Practices: Considering the Use of Technology, Artificial Intelligence, and Process Data for Assessment in the 21st Century. As an early career researcher Network across Europe from different educational systems, we have observed that Artificial Intelligence (AI) is rising across countries. In that sense, the discussion will explore the advantages and disadvantages of AI in assessment research. We look to explore the potential of integrating AI into research while also addressing issues such as the misuse of AI, ethics, fairness, transparency, accessibility, and bias. Also, we will discuss the implications of aligning AI implementation with ethical considerations, guided by the question: Is artificial intelligence considered a helpful ally or a potential antagonist in the field of assessment research?

Discussion Group 2
Room: Leda (n=60)

13:30 **There is clear evidence of inflation in assessment outcomes in many contexts and countries over the past 20 years: Does this matter?**
Isabel Nisbet, Mary Richardson, Stuart Shaw, Lesley Wiseman

Abstract: Grade inflation is a global phenomenon. It is commonly understood as attainment of higher grades independent of actual levels of improved or increased academic attainment. In this Discussion Group, we pose five questions to interrogate whether grade inflation matters, and if so, why it matters and to whom? We offer some specific examples of grade inflation to set the scene, but it is our expectation that participants will feel encouraged to share their broader experiences, and their ideas, within the discussion. The session will comprise four parts: an introduction; contextual examples from stakeholders in education, group discussion finally, a plenary with open discussions and contributions made in the room. We intend this Discussion Group to provide an opportunity for participants to explore in some depth their own thinking on this issue so that dialogue will be informed and stimulated by contributions from all attendees, especially those bringing experiences of different cultures and systems, potential impacts, solutions and maybe those avoided. We anticipate a range of answers to the questions posed revealing the complex picture across Europe, and we hope that participants will leave with new ideas about this phenomenon.

Discussion Group 3
Room: Athena (n=60)

13:30 **Building holistic systems for educational improvement: From curriculum to pedagogy to assessment principles**
Irenka Suto, Carolyn Hutchinson, Tim Oates, Gulbakhyt Sultanova, Stuart Shaw

Abstract: Holistic education is rooted in European pedagogical philosophies. It emphasizes nurturing students' growth beyond academic knowledge, to include social, emotional, and psychomotor skills. In this discussion group we reflect upon what it takes to build holistic education and assessment systems which potentially lead to improvements in outcomes and life chances. Thinking around the ‘golden triad’ of curriculum, pedagogy and assessment is shared through three short presentations. Our presenters on these interconnected areas pose key questions for participants to discuss. We argue that the curriculum should promote interconnected competencies while avoiding overcrowding. Effective pedagogy involves supporting teachers with materials and professional development to enhance engagement and learning. Holistic assessment aims to evaluate interrelated competencies in a construct-valid way, supporting educational improvement at different levels. Collaboration among learning communities is crucial for adaptable assessments and transparent reporting. An overarching question is that of what principles should guide the development of holistic education systems to promote effective pedagogy, prevent curriculum overcrowding, and ensure holistic assessment supports effective learning in educational settings. The discussion group will also explore the challenges in implementing a coherent holistic system, and potential directions for further analysis and research within AEA-Europe's Holistic Assessment Special Interest Group.

Discussion Group 4
Room: Aphrodite (n=50)

13:30 **One size doesn't fit all: How to consider the equity and fairness of access arrangements as we move to digital modes of delivery**
Emma Crampton, Ellen Barrow, Irene Custodio, Meredith Reeve

Abstract: As we transition further towards digital modes of high-stakes assessments, there are opportunities to consider current provision and to think about the different types of access arrangements that may be required to support learner's needs.? Interesting questions emerge about equity and fairness and whether access arrangements level the 'playing field' or may present an unfair advantage to some students. In this discussion group, we explore questions around current access arrangements in the UK and how they may need to change as we move towards more digital modes of delivery: - How do we work together as an assessment community to create clear guidance, that is equitable and fair for all students across formats? - As technology develops, how assessment organisations work together to agree a common approach to how technology updates affect what is available to students as part of normal ways of working and to inform access arrangements? - Should all students have access to all the same accessibility tools? We hope to build participants' awareness of the different aspects of equity and fairness that need to be considered when trying to address and balance regulatory requirements, access arrangements and students needs in digital high-stakes assessments.

Discussion Group 5
Room: Zeus (n=30)

13:30 **Crossing the line: Where did the digital assessment revolution go?**
Rebecca Hamer, Caroline Jongkamp, Rebecca Chivers

Abstract: While the range of assessment items and approaches has increased over the four decades since the first digital assessments, the broad adoption of digital or e-assessment has not kept pace with the technological advances, demonstrating that technology is not the main barrier to its broad implementation. In 2018, AEA-Europe attendees shared their concerns and the barriers to introducing e-assessment more broadly. These included practical concerns regarding technical infrastructure, digital skills of staff and students, stakeholder support and recognition, as well as resources. Two years later, the COVID-19 pandemic led to unprecedented shifts in human interactions, one of which was the worldwide closure of educational institutes. Education and assessment moved online, with institutes introducing infrastructure and tools supporting remote learning and teaching. High-stakes assessment migrated onto digital platforms. Many felt this experience broke through concerns and perceived barriers, leading to full acceptance of e-assessment. However, this revolution now seems to have come to a grinding halt. Using the crossing-the-line format, participants will discuss the reasons behind the return to the "old new" in education. They will choose sides on a series of provocative statements referring the issues identified in 2018 and elaborate their positions to deepen understanding of the enduring barriers.

Discussion Group 6
Room: Christian Barnard (n=200)

13:30 **The assessment impact of AI and how to justify the budget**
Helen Claydon, *Ben Stafford*

Abstract: Assessment has rarely been at the forefront of technological change, but the global onset of Artificial Intelligence (AI) cannot be ignored and the opportunities offered by AI for assessment warrant serious consideration. As part of building any case for change, it should be recognised that a perceived benefit in one area of assessment may not be seen as universally beneficial when considering the upstream or downstream impacts, or the confidence of stakeholders. This discussion group will use a small group thought sharing approach. It will work in two parts, looking at how to evaluate the impact of introducing an AI-driven change to an assessment model, and then how to influence key stakeholders on the value of the change, particularly when this involves securing budget to turn a good idea into reality. The workshop is offered by members of the AEA-E eAssessment Special Interest Group, and will be a launch for further work undertaken by the eAssessment SIG, which delegates may be interested in participating in.

13:30 - 15:30 Poster Session I - 90 second pitches - Presenters commit to stand at posters during Coffee Break
Chair: Cor Sluijter
Room: Akamas A & B (n=550)

Reduced grading in upper secondary school: Exploring students' perceptions
Dan-Anders Normann, Lise Vikan Sandvik, Oddveig Storstad

Abstract: Across educational settings, the assessment methods are widening beyond traditional grading practices. This study examines the perceptions of upper secondary students in Norway regarding reduced grading, a pivot away from conventional assessment methods, within this broader context (Normann et al. 2023). Drawing on a dataset from 1511 students across seven upper secondary schools in Norway, we utilize a quantitative research design (Creswell 2014) to dissect and understand student perspectives on reduced grading practices. While acknowledging the historic durability and continued relevance of traditional assessments, this study seeks to uncover the nuanced impacts of grading on student learning (Koenka et al. 2021; Lipnevich et al. 2021). Prior research (Sandvik et al., 2024) indicated significant differences in how teachers and students perceive grading practices in Norway. To add depth to these findings, our study seeks to map out an explanatory model that clarifies students' attitudes towards these changes. This model will aid in understanding the broader implications of moving away from established grading norms and toward more nuanced assessment methods. We aim to provide insights that could influence policy and shape future educational practices, aiming for a balance that respects tradition yet embraces innovation, to truly capture students' competencies and problem-solving abilities.

Let's Chat! Integrating ChatGPT in student assignments to enhance critical analysis
Chloe Antoniou, Danagra G Ikossi

Abstract: Innovation in Artificial Intelligence (AI) is outpacing changes in pedagogical methodology. It is therefore imperative that we devise creative and out-of-the-box assignments to teach the ethical, effective and safe use of AI, both within education, and as applies to the care of patients. As part of a comprehensive longitudinal digital health curriculum, the Graduate-entry MD Program at the University of Nicosia Medical School has piloted an innovative project which incorporates the popular AI tool, ChatGPT. The project is structured with four main goals, each contributing to the final assessment mark: 1. the guided use of ChatGPT to research a chosen topic, 2. the critical evaluation of the ChatGPT output for factual correctness, strengths and limitations, quality of references provided and ethical concerns that may arise, 3. the creation of an original manuscript using the output of the AI exercise combined with independent research, resulting in a novel, factually sound report on the selected topic, and finally 4. self-reflection on lessons learned and the effect on students' future use of such tools. The consensus student feedback for the assignment was overall positive, highlighting increased understanding of the strengths and limitations of the AI tool which will guide their future practices.

To love or to loathe? Teacher enthusiasm for data use in schools
Christopher Vincent, Katy Finch

Abstract: This study investigated teacher enthusiasm for the use of assessment data in schools in England. The project focused on the views of teachers of STEM (n=35) and non-STEM (n=27) subjects in teaching and middle-leadership roles. The aim of the study was to investigate stakeholder concerns and build on findings in the literature regarding non-STEM teachers' resistance to data use. Q-methodology was used to explore and compare social perspectives on data use between the STEM and non-STEM groups. Results showed there were distinct perspectives within the sample, with teachers who were enthusiastic and unenthusiastic in each group. Enthusiastic teachers, regardless of their subject specialism, viewed data as integral to supporting their teaching, while unenthusiastic teachers raised concerns about workload and data interpretation. Notably, STEM teachers' negative views centred on workload, while non-STEM teachers' reservations were due to lack of confidence in data literacy and concerns about data collection. The findings underscore the importance of tailoring data tools to educators' diverse needs and avoiding assumptions about teacher perceptions based on subject specialism or role. Further research is recommended to investigate the impact of having a STEM background on teachers' use and interpretation of data.

Leveraging Anomaly Detection for Exam Session Monitoring
Mkululi Wami, Antony Furlong

Abstract: Maintaining the integrity and fairness of high-stakes examinations in pre-university educational settings is crucial. Anomaly detection emerges as a promising tool for identifying irregularities during exam sessions, such as aberrant marking. Statistical and machine learning algorithms offer the potential to enhance the reliability and efficiency of assessments by analyzing individual student exam marks to detect deviations from expected patterns. As a result, anomalies such as unusual or questionable exam marks can be quickly identified and resolved, ensuring fairness and credibility of assessments. This poster will present ongoing research from the International Baccalaureate, exploring various data-driven techniques to detect anomalies in students' marks. Using historical datasets, a comprehensive evaluation of different anomaly detection methods will be conducted. The performance of these methods will be compared using metrics such as precision, recall, confusion matrix, and area under the ROC Curve. Based on these results, trade-offs will be considered between different techniques and their complexity versus performance for identifying anomalies. The study will highlight potential biases in the data or approaches and recommend ways to integrate these techniques into live exam settings. This will enhance understanding of students' performance, identify significant outliers that might warrant further investigation, and ultimately improve the assessment process.

Beyond pen and paper: correcting handwriting recognition in subject-specific contexts
Victoria Tassie

Abstract: As most high-stakes assessments in the UK are paper based, student responses need to be digitised using handwritten text recognition (HTR) to apply any natural language processing (NLP) tools. Due to the nature of handwriting, the outputs of HTR are known to be 'noisy' and erroneous. This can lead to less meaningful findings when these outputs are used in NLP tools. We investigated methods of cleaning the outputs of an off-the-shelf HTR tool applied to 17,000 handwritten responses for a short-response GCSE Biology item. We initially curated a subject-specific dictionary to identify words that had been incorrectly transcribed by HTR and words that had been misspelt by students. Over 10,000 responses had an unrecognised word and over 13,000 unrecognised words were found in total. We attempted to correct these words using edit distance to find appropriate alternatives and a masked language model to identify the most probable alternative. The result was a significantly reduced number of unrecognised words in the transcribed responses and improved accuracy of the tools applied to the digitised responses. This method allows for AI-driven tools to be applied to handwritten responses, without the need for human transcribers nor the need to train a new HTR engine.

Exploring Novel Assessment Modalities: The Assessment of Emotional Intelligence within Collaborative Problem-Solving Environments
Deirdre Dennehy, Deirdre Dennehy

Abstract: The assessment of transversal skills has become a pertinent issue for educational and workplace settings as these skills are latent and difficult to measure objectively. Many test developers and educational contexts remain reliant on text-based assessments which often assess an individual's knowledge of the skill, rather than it's authentic use in an educational or workplace setting. This presentation will address this issue. Based on doctoral research funded by Prometric, an account will be given of the case study design employed and the research question explored-that is the extent to which Emotional Intelligence (EI) can be observed and assessed within a Collaborative-Problem Solving (CPS) context. The presentation will detail the manner in which student participants were videoed while working collaboratively on robotics tasks, using an assessment framework designed and adapted for the purposes of the study. Findings indicate that although an individual's EI can be observed and assessed while working on CPS tasks, factors such as the nature of the task and group dynamics have an impact on the types of EI behaviours that may be displayed by participants. This presentation concludes by considering the implications of this research for test developers, policy-makers, researchers and educational professionals.

Exploring the relationships between Extramural English and English Reading Comprehension among 2023 SweSAT test-takers
Teodora Neagu

Abstract: The purpose of the study was to explore the relationship between Extramural English (EE) and English reading comprehension, focusing also on exploring possible gender differences with regard to these variables and the relationship between them. Data was collected in the context of the Swedish Scholastic Test (SweSAT) and included performance data and questionnaire data from 6,079 participants who took the test for the first time in spring of 2023. A questionnaire was developed with items primarily asking for the use of English outside academia, such as speaking, watching, reading, and listening, and with a special emphasis on digital games, that is, whether the respondent plays digital games or not, frequency of gaming, type of game and interactions in English when playing etc. Quantitative analyses were conducted in RStudio to explore relationships between these variables and SweSAT English test performance for males and females, and thus provide new information regarding the role of EE, especially digital games, that has on English reading comprehension among men and women. The preliminary results provide insights into relationships between EE, gender and English language proficiency, leading to educational implications of second language acquisitions outside the traditional educational settings.

The peculiar predictive power of mathematics assessments
Andrew Lyth

Abstract: Baseline assessments provide valuable insights into students' academic potential, helping inform decisions around grouping and additional support. We explored the predictive validity of curriculum-agnostic assessments of students' abilities in Vocabulary, Mathematics, Non-verbal reasoning and Skills. 50,000 UK state school students' results from these assessments were matched with their GCSE results five years later by the UK National Pupil Database team. Using this matched dataset, we examined the relationships between baseline assessment scores and performance in different GCSE subjects, by applying Pearson correlations, regression models (OLS and multi-level) with multiple predictors and the GCSE grade as the outcome. A surprising result was the strength of the relationships between the mathematics score and performances across a range of GCSE subjects, particularly those subjects with no mathematics content whatsoever such as English. Additionally, we explored whether cluster analysis can consistently group some GCSE subjects together based on their relationships with baseline assessment scores.

Onscreen Functional Skills: Insights from a decade of delivery of English and mathematics assessments
Hayley Dalton, Jagdeep Kaur

Abstract: Functional skills qualifications in English and Mathematics (FSQs) are taken in schools, colleges, private training providers and other settings in England. Designed primarily for learners that failed to achieve a 'good' pass in English and mathematics at school, FSQs are intended to be an 'accessible and practical route for students who want to develop and improve their skills.' Using data from a large awarding organisation in England with around two-thirds of the market, this poster looks at the journey from the introduction of FSQs to the present day. We reflect on the shifting demand for paper and onscreen tests across the various learner and centre characteristics. The introduction of onscreen assessment has allowed centres and learners to personalise their assessment experience to a degree. We look at the trends and patterns of the centres and learners who are taking FSQs and whether that can provide valuable insights into the demands and needs going forward for onscreen assessments more broadly in the UK educational landscape.

Assessing student mastery levels in tracked education using diagnostic classification models
Lientje Maas

Abstract: Diagnostic assessment can help determine students' learning needs. Information about skill mastery can be obtained using diagnostic classification models (DCMs; Rupp, Templin & Henson, 2010). DCMs differ from item response theory in that instead of estimating ability on a continuous latent scale, students are classified in terms of mastery of discrete latent variables (i.e., attributes). This removes the need for time-consuming standard-setting procedures and can benefit reliability (Templin & Bradshaw, 2013). This study compares two approaches to model polytomous attributes to enable application of DCMs in tracked education systems. In tracked systems, students can master attributes at the level of a certain track, i.e., there are varying mastery levels. Attributes in DCM applications are often dichotomously defined (mastery vs. nonmastery), yet polytomous extensions have been proposed. In conventional polytomous attributes, attribute levels and their meanings are derived after fitting the data (Bao & Bradshaw, 2019). Alternatively, attribute levels can be substantively defined prior to the data-fitting process (Chen & de la Torre, 2013). We compare these approaches using data from standardized arithmetic/math assessments in five tracks in Dutch secondary education. By evaluating model fit and reliability, the utility of both approaches for applications in tracked education systems is assessed.

The Opportunities of Natural Language Processing for Assessing Essays with Comparative Judgment
Michiel De Vrindt, Anaïs Tack, Renske Bouwer, Marije Lesterhuis, Wim Van Den Noortgate

Abstract: Comparative judgment (CJ) is a method commonly used for assessing essay quality, where assessors compare pairs of essays to determine which is superior in quality. Psychometric techniques are applied to convert the preferences to a quality score for each essay. Although CJ yields reliable and valid scores, wide-spread implementation in educational practice is currently challenged by its inefficiency and limited feedback capabilities. In this study, we explore the potential of using Natural Language Processing (NLP) to address these challenges. We identify the at different stages of the CJ process and discuss opportunities of NLP to enhance the efficiency and transparency in these stages along with potential hurdles. For instance, to alleviate the cold start of CJ, NLP could predict initial essay quality, reducing the number of pairwise comparisons needed for reliable outcomes by 30%. During the assessment, smart pair selection, guided by NLP, could further increase the reliability of the scores, while ensuring that the pairwise comparisons do not become too difficult for assessors. After the assessment, NLP could enable automated feedback, helping to understand how assessors arrived at their judgments and explaining the scores. Ultimately, this three-step integration of NLP in CJ could enhance its efficiency and transparency.

E-assessment of Children's Social Skills
Anne-Mai Meesak, Astra Schults

Abstract: E-assessment can be used successfully to assess young children's skills (Adkins, 2021; OECD, 2020). Next to academic skills, researchers have stressed the importance of social-emotional and self-regulation skills (Day et al., 2019; OECD, 2020; Pisani et al., 2018). Previous research has shown that teachers might not accurately assess all areas (Begeny & Buchanan, 2010; Mashburn & Henry, 2004), paving the way for standardized tests. Teachers in Estonia can freely use an e-assessment instrument, which includes three standardized, norm-referenced and computer-assessed tests for assessing five-year-old children's skills. While two of the tests have been validated using teachers' evaluations (Meesak et al., 2022), the test for social skills has not. The social skills test includes 25 items and 14 short video animations with social situations. A study was carried out in the spring of 2024, where children solved the test with individual guidance from teachers and teachers filled out a questionnaire regarding children's social skills. The children's direct assessment results will be compared with teachers' ratings and the results of the study will be used to validate the results of the test. The poster will present the results of the study alongside examples of the test.

Student participation in developing formative assessment
Monique Dijks

Abstract: Current developments in society and education ask for changes in assessment procedures that are used in educational practice. With greater focus on 21st century skills, critical citizenship became more important over the last decades (Geisinger, 2016). Moreover, students in vocational education desire to be more involved in the development of their education. Student participation in the development of assessments leads to a higher level of involvement of students with the subject matter and may even cause deep learning (Doyle et al., 2019). Assessment then becomes part of the learning process and may contribute to a sense of inclusivity and more appreciation of diversity (Tai et al., 2021). In the current research, 66 students actively participated in the formative assessment process. Literature research was used to inform the process of student participation. A focus group and questionnaire showed slightly positive results of student participation. Students felt more appreciated and had more confidence in themselves. 63% of the students reported that they would like to participate more often. We recommend including students in the development of their assessment in a sustainable and diverse way. Showing students appreciation for their input and visibly using their input is important for students to feel valued.

Challenged by generative AI, assessment practices in upper secondary in Norway: Networking professional assessment competence in the age of algorithms
Mari Bjørnsdotter Vinjar, Thomas Fjeldvik Peterson, Øystein Gilje

Abstract: Summative assessment practices vary considerably across different education systems and policy contexts. In Norwegian secondary schools, approximately 80% of students' final grades are awarded by teachers based on students' performance in the classroom (Fjørtoft, 2020; Hopfenbeck et al., 2013). Consequently, there is a need for professional communities to work on guidelines and build competence in designing assessment criteria and implementing tasks. Over five years, The University of Oslo has been working with a competence development program in partnership with the Oslo Education Agency. A network of leaders and teachers in upper secondary schools is an arena to develop an interpretation community between schools, thereby increasing quality and validity in formative and summative assessments. This presentation highlights two critical questions in this ongoing work: 1. How do teachers work with assessment criteria in specific tasks based upon a competence-based curriculum (LK-20)? 2. How do teachers manage dilemmas that occure when they move between formative and summative assessment? 3. How does the emergence of AI play into the teacher's assessment practices? Building on the expansive circle of learning (Engeström, 2015), the network will develop, implement, and evaluate a manageable and sustainable model for the teachers.

Participatory research in exploring more fair and just assessment in higher education: How assessment can be developed to more equitable learning for students with disabilities
Karina Dyliaeva

Abstract: This poster will outline my PhD study which aims to contribute to critical assessment studies that examine sociocultural processes and assessment practice. Assessment, it is argued, is a key determinant in the systemic exclusion of students with disabilities in higher education. Despite increased interest in student voice and agency, research studies still fail to integrate the diverse knowledge and ideas that students with disabilities may bring to assessment situations and events. My overarching research question is: What role does assessment play in creating barriers to a fully inclusive learning environment for students with disabilities? I will outline the study, how it draws on theories of epistemic injustice, and fair/just assessment to analyse how assessment practices may or may not discriminate against disadvantaged students. The purpose of the study is to gain a better understanding of how educational equity and justice might be implemented in assessment practices for students with disabilities, and whether formative assessment can help achieve this goal. The methodology is a participatory research design, rarely enacted in assessment research, that aims to forefront the social structure of assessment that underlies student experiences as well as attempt to identify student-led ideas for more equitable and inclusive assessment practices.

Stepping stones towards explainable AI marking: extracting keywords and phrases
Alex Dunhill

Abstract: Explainable AI is one of the primary ultimate goals of current AI research. Nowhere is this more important than in AI marking systems, whether it be to ensure transparency in high-stakes exam marking or to provide relevant formative feedback to students. One potential approach is to try to understand and leverage the internal representations that AI models use, so we can probe their 'reasoning'. The internal attention mechanism of transformer models, where they assign numerical values to relationships between words, has strong viability. It has been widely studied in contexts such as image recognition and text generation, to varying degrees of success. Our initial results show that by using the attention values in a language model trained to predict marks, we can also extract keywords and phrases from student exam responses, without reference to an explicit mark scheme. This may represent a vital stepping stone to ensuring an independent AI marker could be made to explain itself.

Enhancing the user experience of Digital Exams: A User-Centric Approach
Mohammad Abbas Abadi

Abstract: The fast-paced development of digital exams led Cambridge University Press and Assessment to reframe its approach, moving to a process of user-centred design focused on carefully understanding different learners' needs, ensuring usability, and paying close attention to accessibility. This approach aims to add meaningful value for learners compared to traditional paper-based assessments and not introduce new barriers or friction for them in this transition. Overall, our user-centred design approach enables Cambridge University Press and Assessment to navigate the digital exam landscape, improving equality of access to quality education. This poster highlights the key stages of the user-centred design process that have been taking place within Cambridge University Press and Assessment for both assessment content and assessment delivery platforms. The approach has been defined as four different stages: - problem discovery, - solution discovery, - design and development, - validation, and iteration. Each stage required applying user experience design methodologies, including user interviews, surveys, storyboarding, wireframing, prototyping, and usability testing in both moderated and unmoderated methods. The poster also underlines the ethics of the Cambridge University Press and Assessment user-centred design process by giving priority to learners' privacy, transparency, informed consent, data protection, and withdrawal rights at every stage.

Examining the Effects of Artificial Intelligence on Secondary school Students' Mathematics Achievement: A Meta-Analysis
Bakyt Alzhanova

Abstract: This study aims to examine the overall effectiveness of AI on higher school students' mathematics achievement using a meta-analysis method. The study findings revealed that AI had a small effect size on higher school students' mathematics achievement. The effect sizes of eight moderating variables, including three research characteristic variables (research type, research design, and sample size) and five opportunity-to-learn variables (mathematics learning topic, intervention duration, AI type, grade level, and organization), were examined.The findings of the study revealed that mathematics learning topic and grade level variables significantly moderate the effect of AI on mathematics achievement. Research questions: 1. Does AI use significantly improve higher school students' mathematics achievement? 2. Which variables moderate the overall effects of AI on higher school students' mathematics achievement? Were examined 35 individual samples of secondary school students and found a positive association between ITS and mathematics (g = 0.35). Each study was examined across five dimensions, including mathematics learning topic (content coverage), intervention duration (content exposure), AI type (instructional resources), grade level (target students), and organization (instructional strategies). Additionally, research characteristics, which might affect the overall effect size, were examined based on the previous metaanalysis. The variable included research type, research design, and sample size.

An experiment into identifying generous and harsh markers using single marked items
Alun Evans, Darren Johns

Abstract: In the UK a large proportion of both A level exams targeted at 18-year-olds and GCSE exams targeted at 16-year-olds are made up of extended response questions. These items are generally seen as a valid form of assessment but marking reliability tends to be lower than when marking for more objective shorter items. This work examines an alternative approach to identifying poor marking which takes advantage of the fact that each candidate is marked by multiple different examiners. The item level data is entered into a Structural Equation Model which allows the estimation of an "expected" mark for each candidate on each item. We then calculate the mean difference between the "expected" mark and the mark awarded for each examiner on each question. If this difference has a large positive value, then the examiner is systematically generous and if it is a large negative value this suggests that the examiner is harsher than expected. This project includes an analysis of how the calculated generosity compared to estimates of generosity from quality control double marking. It will also report on a proof of concept trial which is planned for Summer 2024 and will use this approach in practice.

Assessment Literacy Enhancement of teachers of less commonly taught languages in COVID-19
Thomais Rousoulioti, Dina Tsagari

Abstract: The aim of this study is to contribute to the understanding and development of teachers' assessment literacy (LAL) (Author 1 2020) of less commonly taught languages (LCTLs). In this context the study investigated the level and development of language assessment knowledge and skills of pre- and in-service teachers of Greek as a second language (L2) who attended an online training course the materials of which were based on a major LAL resource (TALE, Erasmsus+, http://taleproject.eu). The study took place within the remote teaching and assessment measures imposed by Covid-19. Quantitative and qualitative data were collected from 89 postgraduate teachers of Greek as L2 via an online pre-/post-survey and from the assignments of the student teachers producted as part of the requirements of their course. The results suggest that even when the group of teachers has common interests and needs in LAL, they have diverse conceptualizations and developmental trajectories of LAL. Overall, the results inform and expand very well-known conceptualisations of LAL in LCTL environments (e.g., Taylor 2009, 2013; Harding and Kremmel 2016) and make research and pedagogical recommendations in developing teachers' language assessment literacy in emergency educational situations.

Exploring multiple summative assessment types for teacher professional development
Berit Haug, Sonja Mork

Abstract: This study explores a multiple summative assessment (MSA) approach with 12 exams across a one-year professional development program (PDP) of 30 ECT for secondary teachers. Factors supporting this approach include: providing assessment forms that offer students ample opportunities to demonstrate their knowledge and competencies; necessity to challenge traditional summative assessment and establish varied assessment forms, especially in the wake of generative artificial intelligence; helping students allocate workload evenly throughout each term, and the grading pass/not pass might reduce stress; and, participating teachers get explicit modelling of MSA to strengthen their assessment practice. 19 teachers followed the PDP integrating natural sciences and didactics. All exams had specific learning goals and corresponding criteria describing expected outcome for pass and not pass. Examples of exams include different types of written, oral and practical tests, and combinations of these, like video-presentation of constructing a musical instrument to assess student understanding of waves. The research question is How do participating teachers reflect on multiple summative assessment forms? Data sources are written reflection notes and interviews, and results indicate that being exposed to MSA raised the participants awareness of their assessment practice and they reflected on whether and how to implement MSA with their own students.

Fostering & Assessing Computational Thinking – Development of a High-Stakes Digital Examination
Abdullah Khan, *Hannah North*

Abstract: In recent years, strides have been made in the development and measurement of computational thinking (CT) alongside advancements in the digital formative assessment of programming. However, gaps persist in this realm, particularly the absence of a high-stakes digital examination that targets the construct of CT. Cambridge is bridging this gap by spearheading the development of a digital secondary school examination for computer science to assess CT. Cambridge is designing an integrated development environment embedded within a secure testing platform, ensuring both the authenticity and security of the examination process. The examination format is designed to resemble a real-world programming experience that include scenario-based questions, with each section focusing on distinct strands of CT, such as abstraction, decomposition, and algorithmic thinking. Additionally, case studies that describe authentic programming problems will be developed as supplementary learning material to aid exam readiness and to have positive washback on teaching and learning. The poster will summarise the work done so far in designing this examination and will also include how perspectives collected from customer research in the form of in-depth teacher interviews and focus groups have been coupled with insights from existing academic research to guide each step of the development process.

15:00 - 15:30 Coffee Break

Foyer outside Akamas Room

Opportunity to visit SIG Banners

15:30 - 17:00 Open Paper Session II

Artificial Intelligence II
Chair: Rebecca Hamer
Room: Akamas A & B (n=550)

15:30 **Maintaining fairness in high-stakes examination marking with AI language models**
Alex Dunhill

Abstract: Ensuring marking consistency between examiners is a challenge for exam boards tasked with maintaining fairness in high-stakes exams. Methods like seeding are necessarily limited as they require additional marking on a small sample of responses. Enhancements (e.g. top-up seeds) have been proposed but they do not fully address these issues. We investigated the potential of an AI language model as a novel marking quality assurance system to address these weaknesses. Using handwriting recognition software to digitise candidate responses, we fine-tuned the language model BERT on a relatively small sample (typically 1,000–2,000 instances) of marking by our senior examiners. This model then predicted a mark for unseen responses and we flagged cases of significant disagreement with the original marker for independent checking. We demonstrated the efficacy of the system on a range of low-mark (≤ 6) tariff questions from GCSE Biology papers and showed that it worked especially well for questions requiring a short-paragraph response. This controlled application of AI in marking leaves human markers in control of how marks are awarded while still acting as to check all non-senior examiner marks, an improvement on methods that use cherry-picked seeds.

16:00 **Using Artificial Intelligence for the Quality Assurance of Examiner Marking**
Darren Johns

Abstract: Quality assurance is important to minimise error with examiner marking. The most used quality assurance processes for itemised assessments at WJEC is 'seed item marking' where, prior to marking live scripts, an examiner will mark a selection of items pre-marked by a senior examiner. In the event of the difference between the two marks exceeding a predetermined tolerance, the examiner is stopped from marking. There are limitations with this method, particularly with cost associated with the setup of seed items. This project aims to train and optimise binary classification machine learning models on quality of marking item level data from the summer 2022 examination series to try to identify the likelihood of marking errors. Of the models analysed the random forest was most successful and had the highest accuracy and precision score. The models were successful in identifying incorrectly marked items but also flagged a significant proportion of correctly marked items as being marked incorrect. The high proportion of false positives mean this model is not suited to replace current quality assurance methods. However, this approach could be used in conjunction with existing methods to target problematic examiners.

16:30 **Can AI write my deepest thoughts?**
Lucianne Zammit, Joseph Giordmaina

Abstract: Recent reforms have heralded a paradigm shift in the national assessment strategy. The Learning Outcomes Framework has shifted national student assessment towards strategies that are more formative in nature. One of the primary formative assessment tools for assessing Ethics in secondary schools is reflective journal writing, which allows students to record and reflect on their personal thoughts, feeling and values related to course content. This makes the learning experience more meaningful to students, helping them internalise ethical concepts. Reflective writing also helps students think critically about ethical issues and moral dilemmas. However, the advent of Generative AI can undermine the validity of reflective journal writing as an assessment tool. It can produce text that mimics human writing styles and thought processes, potentially allowing students to generate journal entries that are not their own original thoughts. This undermines the authenticity of the assignment, making it difficult for educators to assess genuine student reflection and growth. We propose that, to address these challenges, teachers might need to adapt assessment strategies, such as incorporating oral defences of journal entries, assigning journal entries as a class-based task, or designing prompts that require more personalized responses that AI would struggle to replicate accurately.

Fairness & Social Justice I
Chair: Deborah Chetcuti
Room: Zeus (n=30)

15:30 **When is it fair to be generous? New qualifications, standard setting and the sawtooth effect**
Tim Stratton

Abstract: Setting and maintaining standards in new qualifications is challenging. Particularly because students' performance in new assessments may be lower in the first few years due to limited teaching resources, unfamiliar material, and students being less well prepared for the assessments. During the reform of qualifications, we refer to this dip in performance, followed by recovery over the next few years as the "sawtooth effect". In England we aim to compensate for this dip by accepting a lower performance standard in the first few years post-reform, in order to be fair to students taking qualifications during these periods of change. However, applying the same principle to entirely new qualifications presents unique challenges. Unlike during reform, there is no prior benchmark for grading standards. Schools also may adopt new qualifications at different times, meaning patterns of improvement may not be consistent. We explore these issues using a case study of GCSE computer science to exemplify some of the above issues. Using the framework developed by Nisbet and Shaw (2020) we discuss the implications of compensating for low performance at different stages of a new qualification, under different conceptions of fairness, while considering the potential trade-offs between fairness and comparability between years.

16:00 **Reforming the reading personalised assessments in Wales**
Matthew Turner, Ben Tylden-Smith, *Andrew Boyle*, Sefa Sahin

Abstract: AlphaPlus leads the consortium responsible for developing the Online Personalised Assessments (OPAs) for Wales designed with a formative purpose, sat on demand at any point in the school year. The earliest of these assessments have been in place since 2018. As consortium lead, one of our key goals is continuous improvement of the system. Over the course of the last academic year feedback was used to improve the design of the procedural numeracy assessments, with a new and improved design being adopted at the start of the 2023-24 academic year. The update was received positively. The Welsh and English reading assessments have been running since 2019 and have a key design complication compared to procedural numeracy with the inclusion of 'friend sets'. As the consortium's adaptive lead, it is our responsibility to consider feedback and use it to improve the assessments, whilst also ensuring that we still provide robust instrument that will deliver reliable results. To this end, we are currently piloting adjusted designs for the readings. In this paper, we seek to reflect upon the tension between 'pure' adaptivity and practitioner perspectives of how an assessment should operate, and detail the process of reaching a compromise between the two.

Assessment Cultures II
Chair: Andrew Watts
Room: Christian Barnard (n=200)

15:30 **Conceptualisation of assessment in education policy documents – school leaders' room for action**
Jorunn Spord Borgen, *Tine Prøitz*

Abstract: The paper contextualize assessment in the tension between political and professional governance and centralized and decentralized governance. Assessment is historically a topic that is part of the school as a social institution and concerns how we can know that the goals for education are achieved and produce the desired results for individuals and society. The paper investigates conceptualizations of assessment, and what kind of expectations and responsibilities are discussed in relation to educational leaders, in key policy documents from of 1990 to 2020, with Norway as an example.

16:00 **Assessment of Oracy at high-stakes national exams in upper-secondary schools across disciplines and assessment cultures**
Anne-Grete Kaldahl, Ove Edvard Hatlevik

Abstract: Considering the development of AI, the oral exam is relevant. Norway has had oral exams since 1883. The objective is to understand how teachers assess and perceive oracy (understood as speaking and listening competency) across various disciplines by surveying teachers about high-stakes oral national exams in upper-secondary school. The research question is: What are the teachers' expectations for good quality oracy across different subjects' assessment cultures? The findings will indicate what the teachers appreciate as good quality oracy construct in their disciplines. The research design is that first schools are recruited, and second contact is made with teachers at the schools who are sent an online survey. The survey results will be analyzed using concepts and ideas about communication and content from rhetorical theory inspired by Aristotle's modes of persuasion. Teachers' expectations for oracy will be measured quantitatively through three modes of persuasion: logos (subject-specific knowledge), ethos (demonstration of character), and pathos (ability to affect the audience emotionally). The analyzes include descriptive descriptions and the use of ANOVA. Each disciplines' construct will be displayed to see if features are consistent or vary across different disciplines. The discussion will highlight how each subject specific discipline have different assessment cultures of oracy

16:30 **Educational assessment a quarter of a century on: lessons learned and the path ahead.**
Isabel Nisbet, Stuart Shaw

Abstract: The presentation will report on a multi-disciplinary project seeking to reconceptualise educational assessment at the approach to the second quarter of the 21st century. Three main lessons have been learned from this review. The first is that context always matters. Context can determine the purposes and practicalities of education and challenge approaches to teaching, learning or assessment which seek to disregard context. Secondly, there is a lack of fit between constructs for assessment deemed relevant to the 21st century and traditional methods of assessment. The third lesson is a challenge to the traditional caution and risk-nervousness of the assessment world. The presentation will then argue that there is an underlying tension between two possible approaches to assessment: the first, labelled “analytic and narrow”, common to much of the psychometric tradition and highly quality-assessed assessments valuing reliability and comparability; and the second, labelled “synthetic and wide” taking context into account at all stages, aiming to provide rich information about each individual. The presenters advocate a shift in emphasis towards the latter approach. Finally the presenters will consider four possible counter-arguments to their approach, concluding with an appeal for work to develop a new kind of theoretical model for assessment.

Assessment of Practical Skills II
Chair: Tim Oates
Room: Hermes (n=30)

15:30 **Teachers' assessment competence: an evidence from evaluation of summative testing tools**
Marta Mikite, Girts Burgmanis, Inese Dudareva, Dace Namsone

Abstract: A teacher's effective use of summative tests in the classroom can greatly improve student learning outcomes. Thus, there is a growing necessity for measures to assess teachers' practices in using summative assessments, including their abilities to create dependable summative testing tools and score students' work accurately. This study examine quality of teachers' summative testing tools used in classroom and scoring practices. To evaluate tests and related evidence submitted by 80 teachers included in study sample, our research team created summative testing tool measure, a checklist protocol with 25 items. The purpose of this checklist was to reveal how well teachers use criteria like validity, reliability, and fairness in their tests. To gauge how well the instrument work, we assessed inter-rater reliability using Krippendorff's alpha, internal consistency coefficients using Cronbach's alpha for each criteria and Friedman test to determine does significant differences existed amongst implementation of criteria in summative testing tools. The results showed that teachers' summative testing and scoring practices are more valid and reliable than fair. Additionally, our findings suggest that our measure of summative testing tools allows for gathering data with an acceptable level of reliability, which should be further improved in subsequent stages of the project.

16:00 **Exploring Novel Assessment Modalities: The Assessment of Emotional Intelligence within Collaborative Problem-Solving Environments**
Deirdre Dennehy, Deirdre Dennehy

Abstract: The assessment of transversal skills has become a pertinent issue for educational and workplace settings as these skills are latent and difficult to measure objectively. Many test developers and educational contexts remain reliant on text-based assessments which often assess an individual's knowledge of the skill, rather than it's authentic use in an educational or workplace setting. This presentation will address this issue. Based on doctoral research funded by Prometric, an account will be given of the case study design employed and the research question explored-that is the extent to which Emotional Intelligence (EI) can be observed and assessed within a Collaborative-Problem Solving (CPS) context. The presentation will detail the manner in which student participants were videoed while working collaboratively on robotics tasks, using an assessment framework designed and adapted for the purposes of the study. Findings indicate that although an individual's EI can be observed and assessed while working on CPS tasks, factors such as the nature of the task and group dynamics have an impact on the types of EI behaviours that may be displayed by participants. This presentation concludes by considering the implications of this research for test developers, policy-makers, researchers and educational professionals.

16:30 **Improving students' writing skills through assessment criteria by means of podcasting**
Madina Yeskeldi, Nurdana Orazbayeva

Abstract: The aim of this action research is to investigate effects of assessment criteria, technology-facilitated collaborative writing and peer assessment in improving students' essay writing in EFL classrooms. A mixed method approach was applied. It included tools such as pre\posttests, online survey, formal interviews. Essay types consisted of "double questions" ,"discuss both views". Students worked in pairs. Each pair discussed an apposing argument in recorded podcast for brainstorming ideas. They uploaded podcasts on the telegram channel. Then, they wrote an essay collaboratively on Google Docs using modified ICGE assessment criteria. Afterwards, learners sent the link to another pair for online assessment and reflective feedback. Final feedback was provided by teacher. While 33 % of students used complex sentences in pretest, it raised to 70 % in posttest. There was wider usage of tenses, conditionals, modal verbs, cleft sentences, relative clauses, passive voice in posttest. There were twice as many academic terms as in pretest. They wrote paragraphs with more clarity due to improvements in cohesion and coherence. 67% of students in experimental group scored 1-1.5 points higher in posttest. In survey improvements in teamwork, self-regulation, confidence, motivation, advancement in paraphrasing skills, and clarity of content, convenience of using technology were mentioned.

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

Chair: Dina Tsagari
Room: Athena (n=60)

15:30 **Influences on the Perceived Significance of Classroom Assessment Dilemmas**
Christopher DeLuca, Andrew Coombs, Danielle LaPointe-McEwan, Nathan Rickey, Michael Holden

Abstract: Classroom assessment is a promising pedagogy for supporting student learning. However, teachers face a range of challenges that create assessment dilemmas, such that teachers now identify assessment as a core challenge to their professional practice. This study examined the influences on how preservice educators (n=246) perceived the significance of persistent assessment dilemmas. Via an online survey, we collected data on preservice teachers' backgrounds, approaches to assessment, influences on assessment practice, assessment-related emotions, mindset, and perceived significance of six core assessment dilemmas (i.e., addressing parental and student orientation towards grades; integrating technology for assessment; managing the amount of assessment data; navigating different views on the purposes of assessment; responding to the emotional impact of classroom assessment practices on teachers, students, and parents; using assessment to support diverse learners). Through stepwise multiple linear regression, significant predictors of the perceived significance of assessment dilemmas included: endorsement of the purposes of assessment, enjoyment of assessment, professional learning opportunities, the influence of parents/caregivers, school policies/initiatives, and large-scale assessment programs. Demographic, motivational, systemic, and personal factors shaped perceptions of the significance of a range of challenges. The results of this study indicate diverse relationships between specific influences and how educators perceive the significance of assessment dilemmas.

16:00 **Progression to post-16 qualifications in England before and after Covid: analysing the diversity of the cohort to inform policy development**
Kate Sully, Nadir Zanini

Abstract: Progression to further stages of education has been a long-standing topic of interest for researchers and policy makers alike because of the implications for fairness and social mobility. For England, the great deal of available research has focused on transition to university or was carried out before the full roll out of the latest reforms to vocational education. The aim of this research was, therefore, to retrieve up-to-date, in-depth evidence on the progression to post-16 education in England, both in terms of qualifications uptake and attainment. In particular, we focused on four alternative routes available to students and analysed the characteristics of the students taking them through multi-level regression modelling. To compare the attainment of students in different routes, we then took a machine learning approach to simulate attainment if all students had taken the same qualifications. We exploited linked administrative data on the 2019 and 2023 cohort to highlight differences over time. Given the policy interest in progression, findings of this research provide useful insights to inform policy in relation to the post-16 educational landscape as well as assessment systems.

16:30 **Formative gradefree assessments for gifted students at talent centers in Norway**
Tony Burner, Bodil Svendsen

Abstract: This study focuses on students' experiences with gradeless formative assessment at the six talent centers in Norway. Most research has focused on formative assessment in areas where students need extra support or reinforcement, leaving a gap when it comes to gifted students. Gifted students have special learning needs, which if not met, can lead to frustration, a loss of self-esteem, boredom, laziness and underachievement. To our knowledge, there is no focus on the gifted students and their experiences of gradeless formative assessment, despite trials with gradeless foramtive assessments in Norway and elsewhere. The research question we pose is "How do students at centers Gifted and Talented in STEM experience gradeless formative assessment?" and expands on a single-case study. Hundred and fifteen students responded an electronic questionnaire with open-ended and close-ended questions items. Gradeless formative assessment allows trial and error, inquiry-based learning and a focus on the learning process rather than the learning product. As the questionnaire still is open and more students may respond, the presentation will report from the preliminary findings and we will discuss the topic at AEA with the aim of supporting innovative approaches to assessment with gifted students in mind.

E-Assessment II
Chair: Helen Claydon
Room: Akamas C (n=200)

15:30 **The Feasibility of Dual On-screen & Paper Provision for Maths Multiple Choice Tests**
Clair Rawlingson

Abstract: Recently, three major UK awarding organisations announced their intention to place high-stakes assessment on-screen and yet the interim Chief Regulator of Ofqual highlighted, in comments to TES, that radical changes are not to be expected as consistency is the regulator's focus at present. It is in this landscape that my research on the dual provision of maths multiple-choice tests, both on-screen and on-paper, finds its place. In the early stages of Covid, questions were asked regarding whether the digitisation of assessment had taken a pandemic-inspired leap. Putting high-stakes assessment on-screen could be seen as a natural consequence of this move, however it could also be argued that this type of digitisation had already made inroads into the traditional testing space. More diverse technological solutions are undoubtedly the future but the impact of any post-pandemic change on the validity and reliability of assessment must be considered and managed. This paper scopes the considerations that feed into the maintenance of these principles in a digitised space by focussing on one of the simplest technological options, the dual provision of maths multiple-choice tests, and it suggests that research into this area is relevant for both current on-screen testing and future technological assessments.

16:00 **Evaluating the value of AI assisted auto-marking in Cambridge's Implementation of the Digital Mocks Service**
Jesse Dvorchak, Sanjay Mistry, Tom Sutch

Abstract: An initial study was conducted to evaluate the ability and accuracy of a third party, AI auto-marking system, when given a range of different item types across content areas. The questions ranged from 6 to 30-mark responses, mainly text entry questions, from subjects such as English as a First Language, English General Paper and Computer Science. Additionally, the more complex responses from Computer science were included to establish how the auto-marker could be trained to deal with atypical programming style responses. Handwritten responses were transcribed to digital prior to entry into the Auto-marker. Test sets of a minimum of 100 exemplar responses per mark point per question, followed by 2000 anonymised and unannotated test responses for each question were used. Analyses compared the auto-marker outputs with those of human markers and compared the processes used to evaluate marking accuracy across different subjects and questions. We will share accuracy measures obtained by the AI Auto-marker and our verification of those measures. Also highlighted will be other challenges faced across the study.

16:30 **Spotting Hidden Patterns in Language: A Window into Proficiency?**
Ana Ulicheva, Sumita Ishaque, Rose Clesham, *Ellen Barrow*

Abstract: Statistical learning is pivotal in language acquisition, enabling the implicit discovery of underlying language structure through experience. While its exact relationship to real-world language ability remains uncertain, recent studies suggest its potential to predict aspects of both L1 and L2 performance. Notably, performance on statistical learning tasks has been correlated with L2 exam scores, indicating its relevance to language proficiency. In this study, we measure statistical learning by assessing L2 English speakers' sensitivity to letter-to-letter and word-to-word regularities. Utilizing a letter identification task and a word prediction task, participants' implicit knowledge of orthographic sequences and word-to-word probabilities is characterised. Sixty university students, including native and non-native English speakers, participated, providing insights into the relationship between implicit pattern recognition, language exposure, and proficiency. By correlating participants' performance with self-reported language experience and proficiency scores, this research aims to illuminate the relevance of implicit measurements in language assessment.

Psychometrics and Test Development I
Chair: Cor Sluijter
Room: Leda (n=60)

15:30 **Equitable Digital Vocabulary Assessment: What Item Formats do We Need to Build a Fair Vocabulary Test?**
Per Henning Uppstad, Bente Rigmor Walgermo, Njål Foldnes

Abstract: In clinical psychology as well as in education, vocabulary tests function as authoritative indicators of language comprehension. Such tests, however, differ extensively in quality regarding validity and fairness. Also, recent rapid development in digital testing has precipitated a need for vocabulary tasks suitable for computerized adaptive testing (CAT). The present study investigates the validity and fairness of four item formats commonly used to measure group based vocabulary knowledge - single synonyms, contextualized synonyms, categorization and morphology (word parts). For all these measures, differential item functioning (DIF) was investigated between 253 L1 and 129 L2 ten-year-old students. A multiple regression further examined the prediction of the four vocabulary item formats for students' word knowledge (as measured by WISC-V). The present study is among the first to show the different functioning of measures of vocabulary knowledge across language groups, and suggests the isolated synonym item format - if carefully built - to function fairly across these groups. The findings a) call for moderation in what tools to use in equitable vocabulary assessment, and b) suggest how to build sound vocabulary item formats, and c) points to the potential effectiveness of the isolated and contextualized synonym format in future computerized adaptive vocabulary tests

16:00 **DIF items effect on the equating transformation depending on different equating methods and different evaluation criteria**
Marie Wiberg, Inga Laukaityte

Abstract: Test score equating is used to make scores from different test forms comparable and it is common to use the nonequivalent group with anchor test design. When constructing standardized tests, we try to avoid differential item functioning (DIF) items. DIF occurs when groups with the same latent ability but from different groups have an unequal probability of answering a given item. If DIF items occur in the anchor test it may impact the equating transformation. The overall aim was to compare two equating methods when we have DIF in the anchor test when using different methods to evaluate them. We used simulated test data and real test data from the Swedish Scholastic Aptitude Test (SweSAT), which is a multiple-choice binary scored test used for college admissions. The simulations were used to examine different conditions such as presence of DIF and different features of the regular and anchor tests. Preliminary results show that DIF affect the equated values, but which method we recommend to use depend on how we evaluate the method. Practical implications and recommendations for how to handle DIF and lower its consequences as well which evaluation measures are useful when DIF appears in standardized achievement tests are given.

16:30 **The Impact of Non-Cognitive Skills on Academic Achievement: Insights from STEM Secondary Schools in Kazakhstan**
Gulbakhyt Sultanova, Nurym Shora

Abstract: Incorporating non-cognitive skills into the assessment process is integral to a holistic evaluation approach, especially in STEM education. This empirical study explores the impact of 26 non-cognitive skills on academic achievement among students in 20 STEM secondary schools across Kazakhstan. These skills are categorized into four domains – "Academic Behaviors", "Emotional Skills", "Social Skills", and "Identity" – within the framework designed for the national STEM schools. This study focuses on Mathematics and English to provide insights into the differential impacts across STEM and non-STEM subjects. Data from 976 teachers and 13,642 students were collected through surveys assessing non-cognitive skills. Findings revealed distinct sets of non-cognitive skills with varying impacts on Mathematics and English performance, highlighting the nuanced nature of these effects. Key findings include the positive influence of "information processing skills" and "task management" in Mathematics and "time management" and "capacity for consistency" in English, while common skills like "growth mindset" and "grit" were consistently associated with positive outcomes in both subjects. However, certain skills like "energy regulation" and "ethical competence" showed strong negative impacts on performance in both domains. The study underscores the importance of integrating non-cognitive skill development into educational practices and curriculum design to foster academic excellence.

National Tests & Examinations I
Chair: Jannette Elwood
Room: Aphrodite A (n50)

15:30 **Is there any evidence of the saw-tooth effect impacting on learner performance where assessments are more skills & vocationally based? Analysing outcomes data overtime across a range of qualifications/assessments**
Rebecca Bagguley, Jagdeep Kaur, Blake Ashworth

Abstract: There is documented evidence and research that in high stakes assessments, such as GCSEs and A levels in England, when qualifications are reformed there is generally a small dip in performance because teachers and students are less familiar with the requirements of the new content and assessments. This is termed the sawtooth effect. However, there is little research into whether a similar effect is observed in vocational qualifications which have a more blended mix of both academic, skills and vocational focus. This piece of research explores whether vocational technical qualifications observe similar dips in learner performance during reform as other qualifications, focusing on internally assessed assessments that have a stronger emphasis on the assessment of skills. The aim of the research is to start to understand the impact current reforms in England in vocational qualifications may have on learner outcomes during the early years of delivery. This is to ensure that students are not disadvantaged to their peers studying A levels or T levels, as many will be competing for University places. Both quantitative unit level outcomes data from historical years are used as well as qualitative evidence from centres/teachers on what they observe.

16:00 **Where to Draw the Line? - The Limits of Technological Adoption in Assessment**
Dario Pirotta, Francois Zammit, Malcolm Micallef, Analise Grixti, John Muscat

Abstract: This presentation will look at the role of technology adoption in the production and delivery of national assessment in the context of the current syllabi. This study will consider a variety of technologies, ranging from commonly used devices to more complex types. The primary research question is 'To what extent can technology adoption be value aligned to the assessment scope of Malta's current national 16+ and 18+ syllabi?'. The aim is to identify types of technological devices that are already in use or envisioned as suitable for assessment purposes. Through a quantitative exercise held with all examiners, this study will survey the outlook on the adoption of technological devices in different subjects and exam levels, in the context of the existing assessment purposes. The research will evaluate the data collected from the respondents in the light of topical literature. The responses provide an outlook on the opportunities, threats, and challenges that arise in using technologies for current assessment practices.

Higher Education & Assessment II
Chair: Damian Murchan
Room: Aphrodite B (n=50)

15:30 **Machine Learning Modelling: Prediction of Mathematics GCSE 2023 results using 2022 Mock exam outcomes**
Dr Sebastian Nastuta

Abstract: This paper presents a novel approach to predicting student performance in summative examinations. It evaluates the extent to which and the accuracy with which future exam results can be predicted using mock results, a common practice in the UK. This research can be crucial for schools, providing valuable assistance for high-stakes exam preparation. Using a supervised Machine Learning approach, we leveraged item-level performance data from GCSE Mathematics Higher Tier June 2022 to predict outcomes in June 2023 for students who took the 2022 papers as mock exams. Machine learning was employed to train and develop several grade prediction models using item-level results from all candidates who sat the Mathematics Higher in June 2022. Between November 2022 and January 2023, 633 candidates took those papers as mock exams. Their performance in this formative assessment was meticulously analysed to predict their future performance in the June 2023 exam series. By comparing the predicted grades with the actual grades for 524 matched candidates, we have robustly concluded that different machine learning algorithms (OLS regression, logistic regression, support vector machine, decision tree, K-Nearest Neighbours (KNN), etc.) provide 80% and 95% accuracy predictions.

17:00 - 17:45 General Assembly
Room: Akamas A & B (n=550)

17:00 - 17:45 PhD Students
Room: Odyssey Bar

18:30 - 20:30 Event for AEA-Europe Fellows, Practitioners & PhD students

Location: Casa Mespilea

Friday, 08 Nov

8:30 - 9:00 Registration

9:00 - 10:30 Open Paper Session III

Artificial Intelligence III
Chair: Stuart Shaw
Room: Akamas C (n=200)

9:00 **Assessing higher order speaking skills using AI and human judgement: How far can we go?**
Rose Clesham, Sarah Hughes

Abstract: This presentation outlines the research and development of two new speaking item types in an AI/human expert, computer-based, high-stakes English language test. Semi-direct or computer-based tests generally do not offer face-to face interaction opportunities; however, it is essential to research new item types over time to maximise construct and face validity, increase items that require spontaneous production and mitigate against the use of gaming strategies that rely on pre-prepared responses. The two new research speaking items were designed and developed to assess mediation and interaction at different levels of complexity. One of them focuses on the improvisational monologic skill of responding to a given situation, the other an even higher order demand of listening to a lengthy dialogic three-way discussion and giving an extended accurate summary. Both item types were trialled on a large scale. Our presentation will demonstrate the research item types, offer test taker exemplars and discuss interesting patterns of differentiated performances. It was also a key element of the research to gain insights and feedback from key stakeholders. This phase of our research will also include feedback from universities and professional bodies in terms of the fitness for purpose and demand of these trial items.

Perspectives of End-users and the General Public on Assessment I
Chair: Paul Newton
Room: Zeus (n=30)

9:00 **Gaining insights and understanding: School and student perspectives of taking onscreen high-stakes assessments**
Ellen Barrow, Irene Custodio, Meredith Reeve

Abstract: Within a UK context, the potential delivery of onscreen high-stakes examinations presents the assessment community with an exciting opportunity to innovate the current educational landscape. Ensuring that the design and development of digital assessments is as valid, fair and fit-for-purpose as possible is multifaceted. One vital aspect is ensuring opportunities to listen to and respond to feedback from students in relation to their assessment experiences. Building on previous research, this paper reports on the assessment experiences of learners in multiple geographies who in 2024, have taken Onscreen International GCSEs across different subjects. Our findings are informed by surveys completed after onscreen practice (mock) and live assessments, focus groups and case studies of schools who have delivered onscreen high-stakes assessments. This supports our iterative approach to gathering data and making improvements to our onscreen assessments and building positive user experiences. We were able to gather feedback on students' interactions and experiences of the assessment as well as how students interact with the digital components of the onscreen platform. It is crucial not to underestimate these perspectives, especially of the students themselves and their test-taking experiences helping to ensure valid, reliable and fit-for-purpose assessments.

9:30 **Assessment Dysmorphia: the shifting shape of learner achievement**
Mary Richardson

Abstract: A single experience all learners share is that of being assessed at some point in their education. However, it is a particular type of assessment, those tests described as 'high stakes' that have the most significant power in shaping learner identities. This paper presents a new way to characterise the power of assessment: Assessment Dysmorphia (AD): an omnipresent need for only the highest measures of success in all educational outcomes and a limited conceptualization of identity in relation to achievements in education. AD will be discussed in terms of how representation of particular assessment outcomes negatively influences the lives and expectations of learners; just as individuals might modify imagery of their body to create the perfect Instagram self, such ideals distort expectations of the academic self: anything less than a top grade is failure. Using visual and text-based representations of high stakes assessments in the daily lives of learners, this presentation will consider the potential of AD as a means to interrogate the pervasiveness of test outcomes and their negative influence on learner identity. The presenter will reconsider assessment policy design and interrogate ways in which our reliance on test results might be reorientated towards assessing learners in meaningful ways.

10:00 **Preparing classrooms for digital exams: understanding the current experiences of teachers and students in England**
Phoebe Surridge, Faye Walker, Katy Finch

Abstract: As we move towards the introduction of digital high-stakes exams in England, understanding how pedagogical practices and modes of delivery in the classroom prepare students for their assessments is increasingly important. To facilitate AQA's implementation of digital GCSE Polish and Italian exams, we have undertaken a series of school-based case studies. The aim of the research is to provide an in-depth understanding of current classroom practices in these subjects to help prepare teachers and students for the transition to digital exams. Through a series of classroom observations (n=12) and focus groups with 21 teachers, we have captured pedagogical insights into how these subjects are delivered and the extent to which technology is available and used in the classroom. Student surveys (n=193) have provided additional insights into how students use Polish and Italian outside of school, including their interaction with technology. The study has highlighted that while access to devices is important for transitioning to digital exams, successfully embedding digital pedagogy into teachers' practice should be another key consideration. Our findings, which will inform the support packages offered to schools, also suggest that different subjects may require different types and levels of support.

National Tests & Examinations II
Chair: Alex Scharaschkin
Room: Christian Barnard (n=200)

9:00 **Exploring detection for AI malpractice and the future of assessment in the AI age**
Tony Leech, Frank Morley, Emily de Groot

Abstract: One of the major ways that generative artificial intelligence tools such as ChatGPT could affect high-stakes assessment is if candidates use them to produce responses to assessment tasks, especially in non-examined assessments (NEA) like coursework. Under existing regulations in England, such use is likely to be malpractice. To what degree can various detection methods determine whether AI tools have been used to answer NEA tasks? We report on multiple strands of research into this question. We investigated the extent to which commercially available detection tools are able to detect AI-written content if the content is unamended from the output of the AI tool, and secondly if it is adjusted by the changing of prompts. We found that in the latter case, such detection tools can be often evaded. We also explored statistical methods for detecting AI use by looking at patterns of candidate performance on NEA and exams. This method provided a way to see impacts of cohort-level effects such as ChatGPT access, but no impacts were seen in 2023. Finally, we discuss the wider implications of the artificial intelligence age for education, including looking towards the potential changes that are likely to be needed for high-stakes assessments, especially NEA.

9:30 **National monitoring for Wales: Squaring the circle to balance validity, reliability and manageability in assessment design in the context of an evolving, process-oriented curriculum.**
Gemma O'Brien, Ben Rockliffe, Andrew Boyle, Ben Tylden-Smith, Dave Mellor, Hayley Limmer

Abstract: The 2022 Curriculum for Wales represents an ambitious shift from previous curriculum policy and presents a bold new vision for curriculum, teaching and learning in Wales. As part of a wider feasibility project commissioned by the Welsh Government in 2023, this study explored the potential assessment options for development of a sample-based national monitoring programme. This research explored the complexities of developing valid assessment approaches at a national, system level. Key findings from the research that guided the development of the assessment options highlighted that the new curriculum brings with it a complex set of interacting challenges for valid assessment design. The outputs of this project provide insight into the complexities of reconciling the tensions between validity, reliability and manageability in large-scale assessment design in the context of an evolving, devolved, process-oriented curriculum.

10:00 **Exploring the comparability of paper-based and computer-based assessment in GCSE Italian and GCSE Polish: a case study**
Handan Lu, Yaw Bimpeh

Abstract: Previous research into mode effects between paper-based and computer-based assessment suggests a mixed picture of comparability. However, existing studies tend to focus on assessment that uses the test-taker's mother tongue; it remains unclear to what extent the insights are applicable to an exam where a second or foreign language is assessed. The current study, which focuses on GCSE Italian and GCSE Polish exams administered via a customised digital platform, aims to address this gap. The study used a mixed-methods approach; an experimental test was conducted, followed by a pre-test survey to collect participants' demographical information and a post-test survey to gather test-taking experiences. A group of Year 11 students took the paper-based test and went on to take the computer-based test over a two-week period (Polish) and a four-week period (Italian). The comparability of scores across modes of test administration was evaluated using paired-t test and profile analysis provides insights into how students' performance changes over mode. Findings suggested that except for the Polish Listening exams, no significant differences in performance were observed between the paper-based and computer-based testing modes. It is essential for educators and policymakers to consider practical implications of mode effects when making decisions about digital assessment.

Assessment Cultures III
Chair: Christoph Schneider
Room: Athena (n=60)

9:00 **Developing formative assessment cultures and practices in Schools of Music and Performing Arts through e-learning**
Vegard Meland, Julianne Hauge

Abstract: This paper focus on how teachers and school leaders from Schools of Music and Performing Arts in Norway develop formative assessment cultures and practices through an organizationally based e-learning course. Formative assessment is a central part of the Curriculum Framework for Schools of Music and Performing Arts – Diversity and deeper understanding, where assessment competence is outlined together with "The goal is to develop a solid and forward-looking assessment culture suitable for the distinctive character of Schools of Music and Performing Arts". To support teachers and school leaders in implementing and acting upon these directives, The Norwegian Council for Schools of Music and Performing Arts and Inland Norway University developed in 2020 an organizationally based e-learning course focusing on formative assessment. This combined qualitative and quantitative study explores different aspects of developing formative assessment cultures and practices through the e-learning initiative at participating schools. The research is guided by three questions: 1. What are the main formative assessment directives in the curriculum framework? 2. How do teachers and school leaders learn and collaborate in the e-learning courses about formative assessment? 3. What are the implications for the assessment cultures and practices at these schools?

9:30 **Students as decision-makers in assessment design for national systems: lessons from research**
Jannette Elwood, *Kay Livingston*

Abstract: Research has recognised the traditional imbalance of power between teachers and students in assessment practices, where students typically have limited participation in the design of assessment. It has also been recognised that the assessment perspectives and experiences of stakeholders, including students, are paramount for understanding the importance, perception, and implementation of assessment. However, despite this recognition there is less focus on the involvement of students in the design of national assessment systems and classroom-based assessment tasks. Research that investigates student participation in assessment, predominantly focuses on formative assessment emphasising how students can be involved in their learning, rather than them actively contributing to the design of tests and/or examinations. In our presentation we will draw on our joint body of work investigating assessment practices and systems, in non-digital and digital contexts to argue for students to be contributing decision-makers in the design and development of assessment. We argue through a whole school and system approach, that it is necessary for schools, assessment bodies and ministries of education to support students by providing regular opportunities to develop as decision-makers in assessment design. This becomes even more necessary as the use of AI and process data for assessment in schools increases rapidly.

Inclusive Assessment
Chair: Irenka Suto
Room: Aphrodite A (n=50)

9:00 **Examining Assessment from an Inclusive Lens: Challenges and Prospects in a Technological Era**
Charalambos Charalambous, Simoni Symeonidou

Abstract: Both assessment and inclusive education are given significant emphasis in several educational systems worldwide, recognizing their role for promoting student learning. Despite the emphasis paid on each of them in isolation, no systematic attempts have been undertaken to bring together these two fields and explore their areas of convergence and divergence, especially given that assessment can either catalyze or impede learners' inclusion. Arguing that attempts to systematically integrate the two fields need to start by examining their underlying principles, in this paper we co-examine the general principles of assessment with those of inclusive education, searching for their common denominators as well as areas in which they appear to deviate. After identifying areas of convergence and divergence in their principles, using the Cypriot educational context as a point of reference, we discuss the challenges that inhere in developing fair and equitable assessments in a centralized system that over-emphasizes standardization and struggles to incorporate differentiation in instruction, let alone assessment. The paper concludes by discussing how technology, in general, and artificial intelligence, in particular, can help in making assessment more inclusive in educational systems that do not embrace all principles of inclusion. We discuss these ideas both at policy and practice levels.

9:30 **The relationship between homework, digital resource and performance in PISA 2022**
Stuart Cadwallader, Jamie Stiff, Jenni Ingram

Abstract: The literature suggests that there is a positive relationship between time spent on homework and assessment performance, but that this relationship is not simple– many factors are likely to be involved in moderating the strength of the association. One factor that may be becoming increasingly relevant is the extent to which learners have access to, and make use of, digital technology for learning activities in the home. As the processes for undertaking and supporting homework become increasingly digitised, there are likely to be ramifications for pedagogy and academic achievement. This must be considered in the context of the 'digital divide', with learners from different backgrounds having differential access to technology and differing experience of using it. This presentation outlines analysis of data from the OECD's Programme for International Student Assessment (PISA) 2022 study. The analysis explores whether access to and use of digital technology for learning at home moderates the relationship between time spent on homework and assessment performance, while accounting for socio-economic status. The focus is on the UK, but comparisons are also made to a number of other OECD countries that took part in PISA.

10:00 **Assistive Technology in National Examinations – The Maltese Experience**
Edward Mazzacano D'Amato, Dario Pirotta, Ramona Vella Vidal

Abstract: Students with conditions/disabilities are becoming more conversant in the use of Assistive Technology (AT). As a result, in recent years AT has become a commonly requested access arrangement in national examinations. The aim of this study is to find out how effective AT is in the realm of Access and whether all students benefit in the same way. The implications on administration is explored to understand the impact on the logistics involved. This study focuses on four types of AT, namely communication devices, the word processor, reader pen and immersive reader. A mix of qualitative and quantitative methods is used to gather data from different participants at different levels of the education system. Participants include students, users of AT who are currently in Years 9 to 11 and educators/administrators who are involved in the logistics, coordination and teaching related to AT. Results show that AT is not a panacea for all. There are challenges at different stages that need to be addressed for AT to be a relevant and effective tool for access.

Formative Assessment II
Chair: Michael Buhagiar
Room: Aphrodite B (n=50)

9:00 **AI powered adaptive formative assessment: Validity and reliability Evaluation**
Yaw Bimpeh

Abstract: This study introduces an AI-driven formative assessment system designed for secondary school mathematics. The system utilizes learning objectives, cognitive mapping, and various factors, such as detailed competency measures, metacognition, and time, to accurately assess each learner's strengths and weaknesses, thereby determining where to focus additional attention. It delivers prompt feedback to students, allowing them to track their progress and pinpoint areas for improvement. The feedback comprises explanations, hints, and customized remedial resources tailored to individual learning needs. Through empirical studies, the validity, reliability, and efficacy of the AI-powered formative assessment system are discussed and evaluated. The findings indicate that the adaptive system effectively engages students and provides consistent insights into their knowledge and abilities. When asked about their test-taking experience, a majority of respondents (78%) inclined to give it a positive rating. Approximately 40% of participants found the test challenging in some aspects. Around 73% of respondents agreed with the platform's identification of their strengths or competencies. Moreover, 75% of respondents stated that a test of this nature would be beneficial for their learning.

e-Assessment III
Chair: Matthew Glanville
Room: Leda (n=60)

9:30 **Clustering Analysis of Cognitive Processes in Mathematics: Insights from eTIMSS PSI Process Data**
Gaël RAFFY, Adrien Fernandez, Franck SALLES, Aurélie LACROIX, David EL RAIS

Abstract: Using TIMSS 2019 Problem Solving Inquiry (PSI) tasks, this research explores how process data derived from the international log database, combined with didactical analysis, can illuminate students' problem-solving strategies and misconceptions. The study centers on one PSI item, the Robot item, using theoretical inferences and process indicators to analyze students' solving strategies. We employ clustering methods on the eTIMSS 2019 PSI log database, incorporating the five plausible values as explanatory variables. Despite the challenges of using plausible values, the study develops an approach to cluster students based on consistent groupings across models. By manipulating didactically meaningful variable selection and plausible values, the robustness of the clusters is assessed. The analysis confirms most didactical hypotheses, identifying student groups based on mathematical strategies used to solve the Robot item. Overall, the research demonstrates how clustering analysis enhances understanding of cognitive processes in mathematics assessment.

10:00 **Exploring the relationship between students' use of digital technologies and their performance in digital PISA 2022 mathematics assessments**
Irene Custodio, Liyuan Liu, Sebastian Nastuta, Grace Grima

Abstract: The relationship between digital skills and achievement in digital assessment continues to be of significant interest to policy makers and educators as digital resources are increasingly used in classrooms and assessments. Previous research has yielded mixed findings regarding the impact of ICT (Information and Communication Technology) use on students' academic performance, suggesting that this relationship may be influenced by various factors including the purposes and quality of ICT use, as well as students' interests, attitudes, confidence, and competencies. This research explores the relationship between different ICT factors and student performance in PISA digital mathematics assessments. Mathematics was chosen specifically as it was the major domain for PISA 2022 and contain items that move beyond 'paper behind glass' representations. We make use of the comprehensive PISA 2022 ICT questionnaire to investigate this relationship. Our methodological approach uses the PISA ICT questionnaire to develop 14 composite constructs that represent various aspects of ICT usage at school and home, as well as students' interest, self-efficacy, and competencies in digital technology. These scales are then utilised in subsequent correlation and multilevel analyses to determine the impact of ICT on students' performance in the digital mathematics items taken by 15-year-old students in England.

Fairness & Social Justice II
Chair: Isabel Nisbet
Room: Hermes (n=30)

9:00 **Why do returning drop-out students in second-chance middle school programs in disadvantaged schools perform better in Uruguayan national standardized tests than general education returnees?**
Maria Seijas, Gimena Castelao, Jennifer Vinas-Forcade

Abstract: Uruguayan students' test results, grade repetition and dropout are highly related to their socioeconomic status and that of their peers. Vocational schools fare worse than general institutions, mainly given their student composition in terms of socioeconomic status and educational trajectories. Recently increased supports for students in disadvantaged schools have reduced dropout rates, but authorities are questioned for lowering the academic bar. Since the achievement gap remains, it is not clear how to best target supports to help disadvantaged students both stay in school and learn. Using administrative registries of the 9th grade students who took the national Aristas reading and mathematics test in 2018 (n=6584), we rebuilt their educational trajectories and grouped them applying sequence and cluster analyses. We then conducted regression analyses to explore how students pre-test trajectories influence their test scores. Results show the type of trajectory followed influences test scores even after considering the individual and institutional socioeconomic contexts, tracking and gender. Returning students' test performance benefits from enrollment in second-chance tracks and disadvantaged schools. While counter-intuitive, these results highlight the role of student support (such as tutoring) in those programs and schools, as well as the importance of considering educational trajectories when analyzing test scores.

9:30 **Standardized Testing and Social Equity: An Evaluation of Recent Changes in Chile's University Admissions**
David Torres Irribarra, María Verónica Santelices

Abstract: Fairness is expected from large scale standardized assessments, especially those with high-stakes consequences such as admissions tests. Chile has a longstanding tradition of relying on large-scale testing for university admissions and during the last three decades has experienced pressure for increased fairness, particularly among different socio-economic groups and secondary school tracks. This study examines the consequences of a broad modification agenda implemented in 2020 including changes to the tests' content, administration and scoring process. Analyses use national data from a period prior (2018-19) and post modifications (2021-22). The evolution of different fairness indicators for these two periods is examined. We will examine how the results relate to changes in test characteristics and in people characteristics. Specifically, we will take into consideration the test population age, school dependency, average income, and average proficiency. Test characteristics such as average item difficulty and dimensionality for both periods will be explored. This detailed examination will further illuminate the impact of the revised testing procedures on the fairness of university admissions in Chile. Empirical evidence of the implications of recent changes is necessary to satisfy the social demand for fairness in large scale assessments with important consequences on students´ lives.

9:00 - 11:00 Poster Session II - 90 second pitches - Presenters commit to stand at posters during Coffee Break
Chair: Cor Sluijter
Room: Akamas A & B (n=550)

Applying a cognitive model of inference to an existing assessment of reading comprehension
Joanne Kiniry

Abstract: In this poster I reflect on the process of applying a cognitive theory of inference generation to an existing paper assessment (the National Assessment of English Reading (NAER) 2014 in Ireland) through analysis of incorrect responses, and the development of a reading comprehension assessment based on this experience (NAER 2021 pilot study). I explore the difficulties and advantages of this approach, and frame it in the context of potential value-added assessment for learning in large-scale assessments. Latent Class Analysis (LCA) was used to perform a secondary analysis of data collected as part of NAER 2014 in Ireland. A model of inference generation was developed using the work of (Freed & Cain, 2016). This model was then used to inform the development of multiple-choice reading comprehension items where the incorrect responses could be attributed to different inference processes. Learnings from this study suggest that there is potential for the retrospective application of a cognitive model to an existing assessment. Furthermore, the application of such a model looking at incorrect responses could be viable in a large-scale assessment. However, this may only be practical for a limited subtest of items or subscales.

Effect of an Analogy-Based Approach of Artificial Intelligence Pedagogy in Secondary school
Bakyt Alzhanova

Abstract: Artificial intelligence (AI) has emerged as a prominent topic in K-12 education recently. However, pedagogical design has remained a major challenge, especially among young learners. Guided by the Zone of Proximal Development theory and AI education research literature, this design-based study proposes an analogy-based pedagogical approach to support AI teaching and learning in secondary school. This pedagogical approach is centered on human–AI comparison, where humans are gradually shifted from an analogue to a contrast to make visible the attributes, mechanisms, and processes of AI. To evaluate its effectiveness, a quasi-experimental study with mixed methods was conducted. The quantitative comparison shows that the participants in the experimental group learning with the analogy-based pedagogical approach significantly outperformed their peers with the conventional direct instructional approach in all three dimensions of AI knowledge, skills, and ethical awareness. Qualitative analyses further reveal its pedagogical benefits, including demystifying AI through relatable and engaging learning, supporting student comprehension and skill mastery, and nurturing critical thinking and attitudes. The analogy-based approach contributes to the field of K-12 AI education with an age-appropriate, child-friendly pedagogical approach. Notably, AI education should prioritize teaching for student understanding, and AI should be recognized as an independent subject with interdisciplinary applications.

Teachers' perceptions about self-assessment: Value and functionality in language education
ANTONIOS VENTOURIS, DIMITRA TSALTA, OLYMPIA BLATSIOTI, Thomais Rousoulioti

Abstract: The student-centered approach emphasizes the importance of self-assessment to enhance students' self-awareness (Siegesmund, 2017) and self-management. Despite concerns about subjectivity and bias, research supports its effectiveness with proper educational support (Rolheiser et al., 2000). This research explores teachers' perceptions and reservations about the validity and reliability of self-assessment in teaching Greek as a second language in multicultural public schools. 224 teachers completed an e-questionnaire and 12 teachers participated in interviews. Artificial intelligence technologies analyzed the qualitative interview data, while the questionnaire data underwent descriptive and exploratory correlation analyses. The comparative analysis of the results from the interviews and answers provided by ChatGPT for similar questions led to the creation of word clouds through the T-Lab, which demonstrated the main meaning indicators that occur most often. A significant percentage of teachers (75%) consider self-assessment as a reliable and valid method of assessment. However, some teachers still express concerns about the practicality, validity, reliability, and readiness of students to adopt self-assessment, explaining its limited implementation without prior education, despite the general agreement that it improves student performance (Panadero et al., 2016, Rousoulioti et al., 2024).

The perceived difference between computer-based and paper-delivered IELTS in Kazakhstan.
Aliya Khasseneyeva

Abstract: The aim of this study was to investigate the perceived disparities between computer-based and paper-delivered IELTS. As the research aims at understanding perceptions, beliefs and attitudes of the participants taking IELTS, an interpretative case study has been selected as the most appropriate research methodology to answer the research questions (Creswell, 2013; Bryman, 2001). The main data collection method was the survey which allows to obtain answers from a big number of people. Another method was the semi-structured interview which was aimed at obtaining in-depth explanations of the answers by the respondents. The preliminary findings suggest that both test formats have their own pros and cons depending on the individual taking the exam. The respondents identified that listening and reading were more challenging in the computer-based test as they were used to use a pencil to highlight the keywords, while writing was easier in terms of typing and correcting the errors. As for speaking, opinions divided: some people found it more comfortable to take it distantly, whilst others complained against the technical glitches which could have affected their results. Overall, the perception depended on the students' familiarity with the test formal, typing proficiency and comfort with technologies.

Framework for externally quality assuring qualifications that are locally relevant, regionally impactful and internationally competitive.
Brent Abrahams, Mia Andersen, Sarah Howie

Abstract: In a globalised world with growing emphasis on the internationalisation and decolonisation of education, enhancing assessment practices and educational policy has become increasingly vital. The framework presented here proposes that, besides upholding quality standards, external quality assurers should consider: i) the intended curriculum as it pertains to societal expectations of the local context, cultural sensitivities, linguistic diversity, social justice, and economic goals; ii) the actual implemented curriculum in schools; and iii) learners attained or achieved curriculum. These considerations are important proponents of the proposed framework towards ensuring that attained school-leaving qualifications are locally relevant, regionally impactful, and internationally competitive. Achieving local relevance involves tailoring the curriculum to address specific cultural, social, and economic needs. Secondly, achieving regional impact involves preparing learners to contribute to regional development and cooperation. Finally, achieving international competitiveness requires qualifications to meet international standards, and be recognised and portable internationally. Drawing on examples from the Stellenbosch University Unit for International Credentialling, the proposed external quality assurance framework illustrates how external quality assurers can facilitate the international portability of high-quality and contextually relevant qualifications. This framework promotes alignment with international standards while catering to local and regional contexts, towards enhancing the quality and recognition of qualifications worldwide.

Opportunities and Challenges of Externally Quality Assuring Africa's First International School-Leaving Assessment
Mia Andersen, Brent Abrahams, Sarah Howie

Abstract: In the dynamic landscape of international education, guaranteeing the validity, reliability, and fairness of school-leaving assessments is essential for ensuring educational equity and facilitating access to higher education globally. However, the contextual relevance of assessments is often overlooked, leading to concerns about equity, the fairness of policy and assessments, and broader social justice issues, including those associated with decolonising education systems in Africa. To address these concerns, robust external quality assurance processes are needed in order to align assessments with the educational goals and values of the local context, cultural norms, linguistic diversity, and students' backgrounds. As the external quality assurer of Africa's first international school-leaving assessment, the Stellenbosch University Unit for International Credentialling has encountered several challenges, including educational policy vacuums, differing qualification recognition requirements, cultural norms and histories, logistical challenges, contextual concerns, linguistic diversity and political instability. Despite these challenges, there are numerous opportunities, such as the ability to offer an affordable alternative qualification that is Africa centred, that provides access to education globally, thereby enhancing educational equity and address social injustices. It is important that external quality assurers address these challenges and leverage these opportunities to make advances in educational assessment practices in the era of decolonisation.

Improving 8 grade students assessment and speaking skills through asynchronous video making.
Gaukhar Sarsenbayeva, *Ainaz Shadkam*, Zukhra Utesheva

Abstract: This study investigates the feasibility and effectiveness of asynchronous video assessment as an alternative approach to evaluate students' speaking skills. Recognizing the limitations associated with traditional speaking examinations, the research seeks to address them by introducing asynchronous video assessment. The study has been conducted within the context of an internationally accredited school involving grade 8 students who demonstrated sufficient command in English and possess moderate digital literacy skills. An experimental group engaged in the task of video making in Term 2, while the control group receiving a similar task in one of the upcoming terms. The consent form was duly signed, and students were informed of the confidentiality measures securing their anonymity throughout the research process. A comparative analysis was conducted between the results of students' performance in traditional assessments and their experiences with the asynchronous video assessment. The findings indicate positive changes in students' attitudes towards examinations, particularly in fostering creativity and engagement. This study contributes to the ongoing discourse on alternative assessment methods, emphasizing the potential of asynchronous video assessment to enhance the assessment of speaking skills and to unlock authentic assessment for students in educational setting.

Predicting Qualification Outcomes from Early Exams
Richard Harris

Abstract: When awarding of spring series exams for Technical Qualifications takes place ahead of awarding summer synoptic assignments, early awarding of a component of a qualification raises challenges for comparable outcomes of that qualification. Comparable outcomes means if the candidature's performance for a subject is similar over time, then the results profile should also be comparable. Maintaining that comparability is challenging when components are awarded at different times. Analysis was conducted using historical data to assess whether it is possible to predict qualification outcomes from spring exam marks. If qualification outcomes could be modelled at the time of spring awarding, a comparable outcomes approach could be used, as adjustments could be made prior to spring results being issued. A range of classification models were built to predict grade outcomes from spring exam raw marks and other variables. While the quality of the predictions varied across qualifications and models, outcomes were often well predicted. The resulting models allow the standard setting panel to compare predicted and previous actual qualification outcomes during live awarding. This study describes the modelling process, the use of model cards to improve transparency and the piloting of predictions during spring awarding in 2024, including early user feedback.

Exploring Teachers' Views on AI's Role in Assessment in Upper Secondary Schools
Harald Eriksen

Abstract: The introduction of generative artificial intelligence (AI) in the autumn of 2022 quickly raised the question of how and to what degree AI would influence assessment practises in the education system. Provided the inherent learning potential in formative assessment, which also has a legislative status in Norway, in addition to the importance of reliable summative assessment for society, teachers' perceptions of these aspects of assessment and how they are related to AI, can make an important difference (Black & Wiliam 2018). The aim of this study was to explore how teachers in Norwegian upper secondary schools perceive the utility of AI for assessment purposes. Through a survey conducted in autumn 2023 (N = 223) we analysed the data using structural equation modelling (SEM) to estimate the strength of the connections between three latent variables, namely formative assessment, summative assessment and AI utility. Findings indicate a clear correspondence between teachers' perceptions of AI-utility and formative assessment (b = .63) but not summative assessment (b = -0.9). There is also a strong connection between the two aspects of assessment. This study provides valuable insights for teachers navigating changes in the assessment field as well as for decision-makers supporting teachers in this process.

Does the use of ChatGPT in online higher education facilitate learning? A study on students' acceptance and the role of instructor support in technology use
Ioulia Televantou, Ioanna Vekiri, James Mackay, Yianna Danidou, Loucas Louca, Marios Vryonides, Louiza Voniati, Christos Kypri

Abstract: This study explores the integration of AI tools, specifically ChatGPT, into the learning experiences of distance learning students at a private university in Cyprus. Utilizing the Technology Acceptance Model (TAM), the research evaluates the impact of perceived instructor support on the usability and ease of use of ChatGPT, as well as students' behavioral intentions to use ChatGPT. It also investigates the role these perceptions play in facilitating students' achievement and engagement. We report the initial findings of a pilot study based on data collected through an online questionnaire distributed among distance-learning adult students who had some sort of experience with ChatGPT. The analysis was conducted using correlational methods and controlled for variables such as students' privacy concerns regarding the use of ChatGPT and personal innovativeness. Despite the relatively small sample size, the direction of all relationships explored aligned with the predictions of the TAM model. Notably, it revealed a significant negative relationship between instructional support and students' privacy concerns about using ChatGPT, suggesting the necessity of proper education about this technology to alleviate concerns. The findings emphasize the importance of instructor awareness and the strategic integration of ChatGPT to potentially boost overall student engagement in the course.

Exploring the Nexus of AI in English Language Classroom-Based Assessment: Implications and Ethical Considerations
Dina Tsagari

Abstract: This poster delves into the intersection of Artificial Intelligence (AI) and English language classroom-based assessment, examining its utilization, implications, and ethical dimensions. The integration of AI technologies in assessment practices has ushered in transformative opportunities for enhancing efficiency, objectivity, and personalized feedback in English language education. Automated essay scoring systems, adaptive language proficiency tests, and speech recognition technologies represent innovative applications of AI in assessing various language skills. These advancements offer promising implications for learners and educators, including enhanced accessibility, scalability, and personalized learning experiences. However, the proliferation of AI-driven assessment tools also raises critical ethical considerations, encompassing data privacy, algorithmic transparency, and the equitable treatment of diverse learner populations. This abstract underscores the imperative for comprehensive guidelines and regulations to ensure the ethical use of AI in assessment, while fostering a balanced approach that integrates technology with human expertise. By navigating the nexus of AI in English language classroom-based assessment, this poster contributes to ongoing discourse on the ethical implementation of AI in English language teaching, advocating for a balanced approach that prioritizes ethical considerations alongside pedagogical innovation.

The challenge of understanding teacher assessment literacy
Hannah Rowe

Abstract: This research explores the assessment literacy of teachers and how the concept is defined and its characteristics identified. Whilst most stakeholders assume that teachers have good levels of assessment literacy, research shows otherwise. This has an impact on assessments being used in the way they are supposed to and is an important dynamic for policy makers to consider. This study made use of the body of research on assessment literacy, which reveals the numerous definitions of the concept, which suggest a difficulty in knowing what it is. This will challenge teachers even further in knowing if they have become assessment literate. It looks at the frameworks of assessment literacy and how they may be used by teachers to identify if they have the characteristics of an assessment literate educator and if not, identifying the areas for development. This reveals that many are too conceptual to be considered practical, but that some offer policy makers important self-report tools which can be used to ensure that teachers have the assessment knowledge and skills needed to be assessment literate. This is both in general and for the context within which they work, as research shows the importance of context in assessment literacy.

Developing AI literacy in Higher Education through structured assessment practices: Outcomes from a repeated measures design
Evdokia Pittas, Marina Rodosthenous-Balafa, Elena C. Papanastasiou

Abstract: The widespread adoption of Large Language Models (LLMs) such as ChatGPT has precipitated a dual-edged discourse within educational institutions, marked by both considerable optimism and pervasive concerns (Kasneci et al., 2023). While the potential of artificial intelligence to enhance educational and assessment practices is immense, a large portion of academic discourse has focused on concerns around employing AI for dishonest practices like cheating, as well as the strategies for detecting such occurrences. However, very little attention has been placed on how to help students from diverse backgrounds develop their AI literacy, which is imperative for future students and citizens. Therefore, the purpose of the study is to describe and evaluate a model of how AI literacy can be adopted in academic settings. The research questions of the current study are: 1) How do student attitudes regarding AI change after being taught how to use AI ethically for their assignments, and how did that differ based on their pre-existing familiarity with LLMs? 2) To what extent does the quality of the assignments submitted by students change after being taught how to use AI ethically for their coursework? 3) In what ways can AI assist students with learning disabilities with their assignments?

Validation of the Greek Version of the Student Survey of Motivational Attitudes toward Data Science (S-SOMADS)
Ioulia Televantou, Maria Meletiou, Yianna Danidou

Abstract: Attitudes play an important role in enhancing students' learning experiences; yet quality tools to measure them, based on a robust theoretical framework, are not yet available in the emerging field of data science. The present study is conducted by academics participating in the DataSETUP Erasmus + European project, aiming to enhance the professional knowledge of student teachers in the field of Data Science Education. This effort is in collaboration with a research team based in America, led by Kerby-Helm, who is participating in a similar project funded by the National Science Foundation. Utilizing Expectancy Value Theory (EVT) as a theoretical foundation, this latter group endeavors to measure various constructs related to students' attitudes toward data science, including expectancy, subjective task values, and perceptions related to data science. This study presents the findings of an initial investigation into the psychometric properties (validity and reliability) of the Student Survey of Motivational Attitudes toward Data Science (S-SOMADS), translated into Greek and administered to distance-learning adult students in a graduate program at a private university in Cyprus. The instruments developed will be used as tools in research aiming to refine curricula and teaching methods to improve student outcomes and engagement in data science.

HP-FOREG: an infrastructure for assessment researchers
Christina Wikstrom, Per-Erik Lyrén, Inga Laukaityte, Hanna Eklöf

Abstract: The SweSAT is a test used in the admission to higher education in Sweden. It was initially introduced in the 1970s and has since been regularly administered approximately twice a year (see eg Oliveri & Wendler, 2020). Each administration features entirely new items, resulting in an expanding item bank and a substantial collection of test scores. This vast reservoir of data holds significant value for education and assessment research, and a recent initiative aims to enhance accessibility of this data for such purposes. This involves the establishment of a research infrastructure, with codebooks describing empirical data, as well as documentation on prior research and historical materials dating back to the 1960s. Collectively, these resources provide a compelling historical overview from the perspective of educational measurement, shedding light not only on the testing program itself but also on shifts in assessment culture in Sweden and broader trends in admissions testing. This presentation includes a description of the research infrastructure HP-FOREG along with a timeline and historical description tracing the evolution of the SweSAT from its inception to the present day. reference: Oliveri, M. & Wendler, C. (2020). Higher Education Admissions Practices: An International Perspective. Cambridge: Cambridge University Press

New digital mapping tests for young students – our experiences
Oksana Kovpanets, Eren Sübül, Henrik Hung Haram, Guri A. Nortvedt, Andreas Pettersen

Abstract: In 2019, the Norwegian educational authorities decided to develop a new generation of mapping tests for grades 1 and 3 that were implemented in 2022. Data from the first three implementations, year 2022, 2023 and 2024 were analyzed and compared across different years using 2PL IRT and more qualitatively oriented classical analyses. This information were used together with experiences from the test development, from cognitive labs and large scale-pilot studies to scrutinize test content and design and to use the advantages of the digital format to assess the knowledge and skills that students must learn and develop in their early years. These efforts have also helped us investigate how to deal with challenges that occurs with a digital format such as technical challenges or challenges caused by too much information on the screen. The same assessment was used all three years. Following analysis, some changes were made to the tests. Changes that could improve the tests were for example: improvement of audio and video instructions, removing of some items and changing of the order of the items, changes related to test navigation. The poster will present some of our experiences and explain why some changes were made.

Utilising Centre Prior Attainment to Predict GCSE Outcomes in 2025
Thomas Smith

Abstract: The comparable outcomes method is utilised in England to maintain standards between years and awarding organisations. Candidate Key-Stage 2 results are a vital component of the process to achieve a predicted grade distribution. This data is unavailable for the years 2020/1, leaving this method unfeasible in 2025/6. A key alternative evidence source is Common Centres distributions. We explore whether there is any viability in combining these evidence sources to produce a robust and equivalently reliable prediction in 2025. The key concept is to categorise centres, aggregating the results within each category and weighting these by the volumes of entries in each category within a live series. We explore three routes to achieve a valid categorisation: Centre type; splitting Centres into performance categories based on prior outcomes; analysing traits of centres to tune various KNN and DBSCAN models to understand if there is a better classification of centres that delivers a more reliable prediction. We compare the resulting outputs to uncover whether this new method could be suitable. The results of this work may provide a clear solution to the problem of having no prior attainment in 2025 but will certainly deliver deeper insight into the wider use of common centres.

10:30 - 11:00 Coffee Break

Foyer outside Akamas Room

Opportunity to visit SIG Banners

11:00 - 12:00 Open Paper Session IV

Other I
Chair: Lesley Wiseman
Room: Leda (n=60)

11:00 **Educational Certification Theory**
Paul Newton

Abstract: My presentation will argue that educational measurement theory provides an inadequate basis for developing policies and practices related to educational certification and, as such, that educational certification professionals lack a guiding theoretical framework. In response, I will develop the outline of a case for Educational Certification Theory. The principle that underpins ECT is that certification operates at the interface between curriculum, pedagogy, and assessment. Because this is fundamental to its nature, it needs to be theorised explicitly. Resolving tensions that naturally arise between these three disciplines is fundamental to effective certification policy and practice. This analysis is intended to rebut the familiar idea that assessment ought to be the servant of the curriculum. An alternative perspective is developed, which argues that qualification design comprises 3 ordered, but iterative, stages: identifying a profile of intended purposes; specifying a proficiency model; and designing an assessment procedure. Only by drawing upon insights from a range of theoretical perspectives at each of these stages can a qualification be designed effectively. There are no masters nor servants. ECT is fundamentally a democratic discipline.

11:30 **Do naturally curious people score better at high-stake university entrance exams?**
Roman Lyach, Matus Kurian, Adam Lalák, Karolina Letochová, Klára Richterová, Ondřej Šteffl

Abstract: Epistemic curiosity is a basic human trait that helps individuals to learn new things and discover new possibilities. This study examines the relationship between epistemic curiosity and performance on high-stakes tests, particularly the National Comparative Exams utilized for university entrance in the Czech Republic. With access to the test results of around 25,000 applicants annually, the study aimed to measure curiosity among young adults aged 18-21 and its correlation with test scores. Various types of curiosity were assessed, including active, social, scientific, nature, and self-curiosity. While higher curiosity was expected to align with higher intelligence and thus higher test scores, the findings revealed only a weak and inconsistent association between epistemic curiosity and exam performance. While curious individuals exhibited greater joyous exploration, other aspects of curiosity did not significantly predict exam scores. Ultimately, the study concludes that individual curiosity alone cannot reliably forecast performance on high-stakes exams like the National Comparative Exams.

Other II
Chair: Amina Afif
Room: Hermes (n=30)

11:00 **Exploring how technology could mitigate errors in assessment materials**
Lucy Howarth, David West

Abstract: The increased demand on subject matter experts during exam seasons provides a strong case for exploring how new technology could quality assure and check assessments, to reduce dependency on these individuals. Possibilities include machine translation for checking text in modern foreign language assessments, copyright checking, and the use of generative AI for checking questions relating to methods and code in computer science. While these tools may have great potential, there are some important limitations of AI. Example include that its capability to generate natural language across many academic subjects carries a risk of falsehoods through hallucination. It is not able to identify any instance where the description of an external source is at fault, or where the wrong source has been assigned for the task. It would need support with particularly specialised vocabulary. It would struggle to compare complicated mathematical equations for their mutual consistency. Understanding the scope of a mathematical task or interpreting diagrams seem beyond reach for now. The quality of assessment tasks remains the responsibility of awarding organisations, and this cannot be delegated to AI. Technology can support the quality control of assessments, but it cannot replace the role of human experts in high-stakes assessment production.

11:30 **Examination of Gender-Related Differential Item Functioning in University Admission Process in the Czech Republic**
Lenka Firtova

Abstract: Scio, a standardised testing company, administers the National Comparative Exams, which are used in the university admission process by approximately one third of Czech faculties. The two most widespread tests are the General Academic Prerequisites and the Social Sciences; in both of them, men consistently outperform women. An unequal gender representation, with more women taking the tests than men, may be a contributing factor, but there may be other factors as well. In order to investigate the observed gender disparity in the results, we have conducted a Differential Item Functioning analysis (DIF), using 18 tests taken by approximately 25 thousand test takers. In the General Academic Prerequisites test, 17% and 46% of the items displayed DIF in the verbal and analytical sections respectively, with about half of those items favouring women. Certain item types (e.g. “Zebra Puzzles”) displayed DIF across virtually all the analyzed questions, while other item types (e.g. reading passages) displayed little to no DIF. In the Social Sciences test, 34% of the items displayed DIF, with psychology favouring women, modern history favouring men, and the remaining topics (sociology, law etc.) showing little to no DIF. Taking these results into account may promote gender equality and reduce disparities.

Other III
Chair: Dan-Anders Normann
Room: Aphrodite B (n=50)

11:00 **Have writing skills been left behind? Understanding current practice in teaching Writing in schools in England and discussing implications for assessment.**
Alistair Hooper, Grace Grima

Abstract: There is agreement in the literature that there is less evidence about teaching and learning of Writing than about Reading. In international studies (PISA and PIRLS) Reading is used as a proxy measure for literacy, and Writing is not included in the assessments. In England, there has been little progress since the 2012 Department for Education research report ‘What is the Research Evidence on Writing?’ which cited a lack of evidence as to why pupils perform less well in Writing in comparison to Reading and other core subjects, and a lack of understanding of the effectiveness of specific interventions with struggling writers. In order to investigate 12 research questions, a mixed-methods approach was adopted, collecting data via a survey of 743 practising teachers from Primary (n=391) and Secondary schools (n=352), and 57 semi-structured interviews with teachers. Key findings included lack of training in writing for teachers, low pupil confidence and motivation for writing, low teacher confidence in teaching and assessing creative writing. We discuss implications for assessment of writing especially in the context of evolving onscreen assessments in different countries and use this forum for participants to share their views on good practices, concerns and evolving needs.

11:30 **Using data to improve the reliability of internally moderated vocational assessments**
Richard Harris

Abstract: In vocational and technical qualifications (VTQ) in England there is widespread use of internal assessment, where work is marked within centres, often by teachers. Awarding organisations use external moderation processes to ensure reliability and fairness of these internal assessments. Understanding how the moderation activities are performing is key to ensuring that assessments are fair, valid, reliable and consistent across subjects, centres and time. This study is part of a range of workstreams at an awarding organisation designed to identify 'the state of their art' of external moderation processes, with the view to drive improvements principally through leveraging and communicating data. Some variation was discovered in moderation between subjects, the underpinning reasons for which were investigated. Quality Delivery colleagues found data-driven insights to be valuable in ongoing improvement activities. In addition to a focus on processes and documentation surrounding internal assessment activities and how data can inform and improve them, technical solutions such as statistical moderation were modelled. This study presents the technical and data-driven findings of the work, seeking to promote research in the VTQ space, in order to solve what are seen as difficult problems in internal assessment.

Process Data I
Chair: Gulbakhyt Sultanova
Room: Christian Barnard (n=200)

11:00 **Purifying the ability from external variables.**
Daniil Talov, Denis Federiakin

Abstract: In social sciences, measuring constructs and evaluating their relationships with other variables is often complicated by external factors. These external variables can bias estimates and provide alternative explanations for results. This creates a demand for quantitative methods that correct estimates of the relationships between variables. The aim of this study is to investigate methods for purifying the target ability, measured by the target test, from external variables. Three methods are used for purifying abilities: (1) linear regression, (2) orthogonal and (3) oblique bifactor models. To investigate the functioning of these methods, we conduct a simulation study. The simulations demonstrate that it is possible to purify the target ability in two ways: by completely removing (via regression and orthogonal bifactor models) or by partially retaining (via oblique bifactor models) the confounding variance. We also provide a real-data example: the assessment of mathematical literacy in the first grade of elementary school. At this age, children experience difficulties with reading, so the tasks have to be voiced. In this case, phonological literacy interferes with the success of solving items. The differences in interpretation of these methods are discussed in application to the real-data example.

11:30 **Prompting ChatGPT for help with crunching and analysing large data**
Gilbert John Zahra, Ramon Grech, Gian Paul Gauci

Abstract: In this study we take an action research approach by directly and swiftly updating our approaches based on the research. We: asked ChatGPT how to address specific tasks related to large datasets; tried this out; compared it to our current approach; returned to ChatGPT if needed; and updated existing procedures. Our main research question is How can ChatGPT help address dataset issues? However, another question emerged along the way: • How is ChatGPT prompted effectively? With new technologies, existing words sometimes take on new meanings. Prompting refers to the way a human communicates with AI, by for example writing concise questions. Questions can be worded differently and answered differently. Similar to the fact that there are best practices in asking questions in examinations, there are also best practices in prompting AI. At times, ChatGPT directed us to terms and processes which we were unaware of, some of which, like scripts, may seem daunting to the 'non-IT' user. While pitfalls in ChatGPT will be discussed (hallucinations, data-poisoning, and non-alignment), our experience has been mostly positive. We envisage that prompt engineering shall be a necessary tool for data processing in educational assessment and beyond.

Summative Assessment
Chair: Catarina Correia
Room: Zeus (n=30)

11:00 **Exploring the Practicality of Adaptive Comparative Judgment as a Summative Assessment Method in Legal Education**
Kjetil Egelandsdal, Eva Hartell, Jan-Ove Færstad

Abstract: This paper presents a study on the practical application of Adaptive Comparative Judgment (ACJ) in assessing student exams within legal education at the University of Bergen, Norway. ACJ is a method used to evaluate and compare the quality of qualitative work, involves assessors judging pairs of items relative to each other. Unlike traditional assessment methods, which rely on absolute judgments, ACJ requires assessors to assess items in pairs, determining relative quality rather than absolute quality. The study examines the feasibility of integrating ACJ into existing educational frameworks, addressing key challenges such as resource allocation, grading transparency, and assessing exams with multiple tasks. Specifically, the research focuses on assessing student exams within an undergraduate course on Property and Intellectual Property Law, with a total of 300 exam papers and eight examiners participating in the study. Utilizing RM Compare as the ACJ software, each assessor was instructed to dedicate 35 hours to comparative judgment, aiming to establish a reliable rank order of the exam papers. The findings reveal several challenges associated with the practical implementation of ACJ, including issues of resource allocation, grading transparency, and the assessment of exams with multiple tasks. Potential solutions to these challenges will be discussed.

11:30 **Examination in the professional sector: towards the use of Linear On the Fly Testing**
Angela Verschoor

Abstract: Flexible, secure, transparent, fair, and reliable are important, but conflicting, properties of large examination and certification programs. When more than 500,000 candidates per year take the same exam, it is difficult to guarantee that the exam items will remain secret for a prolonged period of time. Thus, the question was if Linear On the Fly Testing could be the solution to a flexible and transparent examination procedure, while maintaining a high level of security, fairness and reliability. As each candidate enters the test center, a linear test form will be assembled in such a way that it is unique while maintaining all adhering to all relevant restrictions. By applying ATA-methods, LOFT can be shown to produce a unique exam form for candidate, virtually without compromising psychometric properties of the test forms, and maintaining a relatively low risk of predictability, and thus breach of security, even after 1,000,000 (simulated) candidates. Currently, the agency introduces LOFT in a 'soft' way, slowly increasing the number of weekly test forms assembled by LOFT until in a few years' time, all candidates will receive a personalized exam form.

E-Assessment IV
Chair: Graham Hudson
Room: Athena (n=60)

11:00 **Digitalising examinations: developing qualifications policy to enhance the validity, engagement and inclusivity of assessments in GCSE qualifications in Wales**
Dean Seabrook, Cassy Taylor

Abstract: On-screen testing has grown in prominence within national qualifications in Wales in the last decade, with an increasing number of learners in full-time and work-based learning taking digital examinations across a wide variety of subjects. Recent qualifications policy changes will see an expansion of on-screen testing into more than a dozen GCSE qualifications, with changes being introduced in September 2025. This paper will outline the theoretical underpinning of the approach taken to developing the relevant policy to ensure that GCSE qualifications will make the best use of the digital technologies available to schools and learners. The paper explains how the intended benefits of aligning examination assessments with learning processes, and enhancing validity, engagement and inclusivity, have shaped the policy. It also acknowledges the influence of manageability considerations on the overarching policy changes. The paper contextualises the policy within a broader programme of modernising qualification assessments in Wales, which is underpinned by mixed-methods research, including semi-structured interviews, focus groups and practical workshops. As the introduction of these digital-only examinations indicates a trend towards digitalisation, rather than digitisation, of assessment, the paper concludes with a discussion of the potential impacts of this approach on future qualification design.

National Tests & Examinations III
Chair: Carolyn Hutchinson
Room: Aphrodite A (n=50)

11:00 **"Trust, but verify": Perspectives of test-takers on validity and trust in a university entrance examination**
Pok Jing (Jane) Ho

Abstract: The title of the study comes from the Russian proverb "doveryaj no proveryaj" (trust, but verify). While it seems to be an oxymoron, the ideas of trusting an exam and evaluating the claims made about the exam go hand in hand in educational assessment. Public trust gives value to the exam results and could (though not necessarily) be strengthened by validation evidence. Since the concept of trust has not been extensively investigated empirically in tandem with validity, this study explores the potential interplay between them. Using the university entrance exam in Hong Kong as a case study, data from semi-structured interviews with test takers are analysed. Given the likely implications of COVID-19, the study also considers the impact of the pandemic on their views. Preliminary findings show a mixed picture of trust in the examination, though it seems largely unrelated to the global health crisis. Political interference has been cited as a major factor that may undermine their trust in the exam board and, by extension, the exam. The availability of operational validation evidence such as assessment frameworks and exam reports appears to be helpful in establishing trust, whereas research evidence seems to have little bearing on their perceptions.

11:30 **Examiners' assessment feedback and announcement of grades to students after summative oral exams**
Marte Søve Syverud

Abstract: This study investigates examiners' practices of giving students feedback and announcing grades after summative oral exams. These exams are carried out by pairs of examiners, one internal (the students' own teacher) and one external (a practicing teacher from another school), who co-assess students after oral exams in the school subject Norwegian language and literature. Grades from these high-stakes exams are included in students' final diplomas and form the basis for rankings for placement in further education; however, the exams are non-standardised and have no national standards or assessment criteria. The following research questions are addressed: What feedback do examiners provide students after completing summative oral exams? How are feedback and grades announced to students? Theoretically, the analysis draws on the conceptual framework of weighting, which differentiates between universal and differential grading. The data include 29 transcribed video recordings of authentic oral exams in one lower and two upper secondary schools in Norway carried out in 2019. Preliminary findings indicate that while oral exams are considered summative, examiners give students advice for future assessments. External examiners normally announce grades to students and provide most of the feedback. The feedback includes both universal and differential aspects, particularly mastery of the oral format.

12:00 - 12:45 Keynote Speech
Chair: Damian Murchan
Room: Akamas A & B (n=550)

Associate Prof. Joshua McGrane: University of Melbourne

12:45 - 13:45 Lunch

Armonia Restaurant

13:45 - 15:15 Open Paper Session V

Artificial Intelligence IV
Chair: Beth Black
Room: Akamas C (n=200)

13:45 **Embedding digitally-mediated formative assessment in the teaching and learning of chemistry: Lessons from International Schools in China.**
Xiaohui Yang, Damian Murchan

Abstract: This study investigates the implementation of formative assessment (FA) in the context of high school chemistry in International Schools in China. It provides an interesting addition to existing research, being situated in a unique, under-researched school system characterised by extensive penetration of digital technology for teaching and learning. The extent to which the affordances of the technology in these schools extend to assessment is of particular interest. The study provides data on chemistry teachers' knowledge, beliefs and application in relation to FA and the variables that shape effective use. A mixed-methods design was used to highlight themes from the literature and to capture the views and practices of twenty teachers working in four international schools in two cities in China. Online questionnaires and remote interviews were employed. Findings indicate that FA is a policy priority in the schools and implementation tends to focus on particular strategies. Digital platforms facilitate implementation of assessment in practice, especially testing, but there is some ambiguity about how teachers differentiate between assessment for formative and summative purposes, with implications for students' effective use of feedback. The results provide interesting insights about the use of FA in a technology-rich school culture.

14:15 **Exploring the Potential and Pedagogical Implications of Pre-instructed AI ChatBots in ESL classrooms.**
Øystein Gilje, Nina Eriksdatter, Trond Ingebretsen

Abstract: During the last year, GDPR-safe AI systems have been implemented as teaching- and learning tools in Norwegian schools (Author1 et al., 2024). However, research findings on their effectiveness are still limited. This presentation focuses on how teachers in the ESL classroom customize chatbots to meet specific curriculum objectives. In a research-practitioners collaboration between the University of Oslo and Oslo Municipality, the study analyzes how these chatbots can provide personalized feedback and offer tailored learning pathways. Preliminary findings indicate that they can serve as dialogue partners and immediate writing feedback providers, encouraging students to explore different language learning strategies and engage with various digital resources. In a multicultural context, they have the potential to promote cultural awareness. However, findings indicate that educators must be mindful of potential biases in the AI models and ensure a culturally diverse approach to fostering students' understanding of the English-speaking world. Although exposure to an American language model may unknowingly promote American norms, teachers can facilitate a broader understanding of different cultures, encouraging open-mindedness and intercultural competence among students. Overall, the findings in the project indicates that AI-driven assessment systems have the potential to significantly enhance English language acquisition and promote effective learning in ESL classrooms.

14:45 **"I used to know but I'm not sure now – what was I made for?" Teachers' concerns about the use of artificial intelligence in classroom assessment.**
Gabriel Cipriano, Isabel Alexandre, Susana da Cruz Martins

Abstract: In the 20th century, teachers' functions and schools' purposes have been the subject of different reflections, perspectives, and tensions among policymakers and academics. In the near future, with the introduction of Artificial Intelligence (AI) in the classroom, the human relationships, pedagogical practices, assessment, and learning processes might be severely wobbled and reconfigured, where individuality will reach levels never seen before. Facing the absence of regulation on the introduction and use of AI in the Portuguese school system, we decided to organise 12 exploratory interviews with teachers, to have a deep comprehensive understanding about the current use of AI in Portuguese classrooms, and the needs and concerns about its use in the future. Preliminary results unveiled that AI is not yet used in Portuguese classrooms. Additionally, some teachers revealed that they do not know how to efficiently use AI in the future, personalised to the specificities of their subject areas and adapted to the age/grade level of their students. The results also point out huge uncertainty and concerns about what the role of AI and the teacher will be, and the need to develop further research for the production of a policy brief to support policymakers with the AI regulation process.

Psychometrics and Test Development II
Chair: Rose Clesham
Room: Athena (n=60)

13:45 **Beyond agreement: Expanding validity evidence for automated essay scoring using contrastive explanation**
Sarah R. Hughes

Abstract: While the vast majority of automated scoring research has hinged on whether humans and machines agree on scores, it is arguably more important to develop methods for determining why humans and machines agree. Evidence that demonstrates humans and machines are influenced by the same aspects of a response when determining a score would provide persuasive support for construct validity. However, modern automated scoring systems rely on complex and opaque models to score responses, and it is not always possible to directly explain how a particular score was determined. New methods are needed to address this gap in validation practice. This presentation reports on the second phase in an overarching project to expand the construct validity evidence for assessments that use automated scoring. The study uses both expert human judgement and eXplainable AI (XAI) techniques to identify influential factors in a scoring decision, manipulate the response to effect a change in score, and test if human scorers and machine scorers are similarly sensitive to the changes. As automated scoring systems become more commonplace in high stakes assessment, the findings of this study will be of interest to researchers and practitioners seeking methods of evaluating construct validity in automated scoring.

14:15 **Population ability estimation and confidence of ability shifts**
Annemarie Timmers, Marieke Van Onna

Abstract: Multiple exam administrations can be equated by fitting an IRT model on the exam data when the design is connected. For proper interpretation, the population abilities need to be translated to familiar scales. In this study, this is done by generating plausible score distributions on the most recent exam version. Obtaining the standard error of the mean plausible scores per population, however, is not trivial. In this study, a computationally intensive bootstrap approach is used. It provides an estimation of the population means per bootstrap sample. These allow for unbiased plotting of the standard errors of the population means, and unbiased estimation of the standard error of shifts in population mean between years. This method is applied to data from over 30 end-of-school subjects in the Netherlands, using the data from 2015 up to 2024. The shift in population ability from 2023 to 2024 is used to equate the exams of 2024.

14:45 **Clustering Subjects Based on Resit Score Improvement**
Heleen de Lange, Marieke Van Onna, Bregtje Seton

Abstract: End-of-school exams in the Netherlands provide a resit opportunity. Between the communication of results of the first attempt and the resit administration, there is time for additional preparation. An improvement in score between the two attempts has three sources: regression to the mean, an effect of learning, and the difference in test form difficulty. In our resit analysis, we attempt to discern these three sources. To accurately estimate differences in test difficulty, we assume consistent learning effects within clusters of subjects. However, the effect of learning may differ over subjects. In this study, a wide range of subjects is clustered with respect to similarity in effect of learning. Homogeneous clusters of subjects, with respect to the effect of learning between the two attempts, are pivotal for fair grading of the resit exams. Data of 44 exam subjects since 2016 are analyzed, leaving out the Covid years 2020, 2021, and 2022. Subjects vary in level (vmbo, havo, and vwo) and in domain (science, language, humanities, and art). Different clustering methods are applied to obtain optimal clustering of subjects.

Test Development I
Chair: Christina Wilkstrom
Room: Leda (n=60)

13:45 **Adapting Innovative Approaches to Enhance Creative Thinking Assessment**
Jonathan Heard, Claire Scoular

Abstract: This paper presents a clear definitional framework for assessing creative thinking and shares the findings from a pilot study aimed at assessing the skill among 10–16-year-old students. Preliminary analysis suggests promising psychometric properties of the items, indicating potential validation of the assessment framework. However, the study also reveals limitations in assessing creative thinking within the confines of existing online assessment platforms, highlighting the need for flexible and adaptable assessment methods. Originally designed for paper-and-pencil administration, the transition of the pilot to an online format proved restrictive from a test development standpoint. The paper discusses inherent limitations of traditional online item types for assessing creative thinking and proposes leveraging process data and other technological advancements to enhance test validity. Exploring these technologies offers exciting prospects to overcome current assessment constraints. The paper presents item examples to demonstrate innovative approaches to enhance items utilising process data into future assessment designs.

14:15 **The impact of mode of assessment on examinee cognitive processes**
Ezekiel Sweiry

Abstract: While most studies on the comparability of equivalent paper-based and digital assessments focus on differences in difficulty, there is a growing body of evidence showing how mode of assessment can affect examinee cognition and strategy. This evidence comes not only from comparability studies, but also from research in areas of cognitive psychology, including how reading and writing processes vary between paper and screen. Research in this area has increasingly utilised technological approaches (e.g. eye tracking and keystroke logging) and focused on exploring how cognition can be impacted by digital device type (e.g. desktop or tablet). Based on an extensive literature review of this evidence, the first part of this presentation considers how mode of assessment affects the mental processes of examinees at each stage of the question answering process, from reading the question through to composing a response. In the second part of the presentation, the implications, beyond those relating to the difficulty of equivalent paper-based and digital assessments, are explored. Consideration is given to how changes in test taker cognition and strategy caused by the mode of assessment also constitute changes to the construct being assessed, and to how we might determine the construct relevance of these changes.

14:45 **What kind of contextualisation is appropriate for assessing application of knowledge? Towards a more comprehensive framework for embedding examination questions in context**
Filio Constantinou

Abstract: Embedding questions in context is a common method of assessing students' ability to apply their knowledge to new situations. However, what constitutes appropriate context remains a controversial issue. This study differentiates between two validity-driven interpretations of context appropriateness in assessment: (a) the extent to which the context allows students to demonstrate their true knowledge and skills, and (b) the extent to which the context is consistent with the specific aims (or claims) of the course/qualification of which the assessment is part. While the former interpretation has been extensively researched, the latter is less – if at all – explored. This study examined this latter interpretation. Specifically, it investigated the extent to which the context used in 527 Functional Mathematics questions was consistent with the aims of the respective qualification. The analysis led to the development of four contextualisation principles: deep contextualization, context balance, context unpredictability, and context purposefulness. This presentation will introduce the four principles and will discuss how they can be used to guide the development and/or evaluation of tests that aim to assess students' application skills. It will then combine the two interpretations of context appropriateness to propose a more comprehensive framework for assessing students' application skills.

E-Assessment V
Chair: Mary Richardson
Room: Zeus (n=30)

13:45 **Establishing modal effect in high stakes assessments: Findings and recommendations for data collection and methodology based on a trial of paper vs onscreen assessment of GCSE English**
Kevin Mason, Sebastian Nastuta

Abstract: In England, we are yet to see a significant transition from paper-based to digital assessments for school-based national assessments for the General Certificate of Secondary Education (GCSE). To corroborate the validity of digital assessments, a comprehensive programme of research is being undertaken to ensure confidence that the move to majority-digital assessment will not introduce new sources of unfairness into the education system. The present study is a part of this programme, examining whether the mode of assessment for GCSE English Language introduces construct-irrelevant variance in assessment outcomes. GCSE English language consists of two assessments. 1734 students across eight schools were recruited and split into two groups. The first group took Paper 1 digitally and Paper 2 on paper; the second group took Paper 1 on paper and Paper 2 digitally. Background information relating to the students was collected. Three analysis methodologies were applied to the data: Ordinary least squares regression, multilevel modelling, and differential item functioning. This paper describes the hurdles that needed to be overcome to obtain such a significant sample of students, and how these challenges impacted on the methodologies that were used, and the confidence in the outcomes and lessons to be learned for future work.

14:15 **Functional Skills Qualifications: Investigating shifts in demand for onscreen and on-demand maths and English assessments in England after over a decade of delivery.**
Hayley Dalton, *Jagdeep Kaur*

Abstract: Using entry data obtained from a large awarding organisation, with around two-thirds of the market for Functional Skills Qualifications (FSQs) in England, we use a mixed method approach to explore how demand for these qualifications has changed over time. With a focus on onscreen and paper-based delivery, we consider how distinct types of providers and students are accessing different assessment modes. Through this analysis, this paper sets out the reasons for the changing demand and asks if we can generalise these findings to inform future demand for onscreen assessment. We explore the data that show a shift in demand across time from younger to older learners and from paper to onscreen and back to paper. On face value the data show a decline in demand for FSQs, but underneath that we use qualitative data to explore further how providers have evolved their offer. Themes explored will include funding, policy drivers, availability of assessment and provider resourcing. In conclusion we consider to what extent we can use these findings to inform future planning for the provision of onscreen assessments for other types of qualifications and high stakes assessments.

14:45 **The Power of Situational Interest in Classroom Reading Assessment - reciprocal relations of interest, self-efficacy, and skill**
Bente Walgermo, Per Henning Uppstad, Njal Foldnes

Abstract: Classroom assessment influence students motivation for further learning. In order to formatively utilize these motivation and skill dynamics; school assessments should mirror engaging classroom instruction and ultimately enhance students' interest and motivation for learning, even while they are being assessed. The present study investigates third grade students' motivation and performance (longitudinally) over 3 sections of a mandatory reading screening test (N=526, mean age 8.6, 49.7% girls). Before and after each test-section, students' situational interest and self-efficacy were reported in addition to willingness to undertake similar tasks in the future (continued interest). A cross-lagged reciprocal effects SEM-models revealed pre-task situational interest as the strongest predictor for continued interest for every subtest, controlling for general reader self-concept and interest for reading. Only for spelling, continued interest was also supported by post task self-efficacy. These results indicate that level of situational interest reported when first introduced to tasks was the strongest predictor for continued interest. Meaning that situational interest predicts continued interest for such reading tasks above and beyond self-efficacy and actual task performance. Consequently, the importance of triggering situational interest in assessment situations for future desire to take part in reading activities, for test design and classroom instructional practices are discussed.

National Tests & Examinations IV
Chair: Louise Badham
Room: Aphrodite A (n=50)

13:45 **Exploring the stability of VA-estimates for school accountability systems using a simulation approach**
Tom Van Ransbeeck, Koen Aesaert, George Leckie, Wim Van Den Noortgate

Abstract: Educational assessment is integral in guiding the understanding of student learning and informing policy. Many educational systems use value-added (VA) modelling to quantify and compare effects schools have on their students' performance. VA is defined as the extent to which a school contributes to its students' learning progress. The stability of school effects/VA-estimates is of paramount importance for intended purposes e.g., school choice or the identification of (in)effective schools. Unstable VA-estimates might be caused by real changes in schools or by characteristics inherent to the statistical model used. This study focusses on one type of characteristics; i.e., sample size, studying the stability of VA-estimates – under the conditions of a joint-VA-model and the traditional approach of using separate VA-models. A simulation study was set up. The results demonstrate the unbiasedness of the stability estimator in three-level-joint-models. Further, a minimum number of students per school seems required to avoid finite-sample bias. In addition, a sufficiently large total sample size permits adequate inference – mostly influenced by number of schools. An overall minimum of 50 schools in combination with a reasonable number of students or a much larger number of schools to permit smaller school sizes are general conclusions.

14:15 **Engaging teachers with how standards are set in high stakes summative assessments: The case of Welsh GCSEs.**
Stuart Cadwallader, *Michelle Meadows*

Abstract: In Wales, a new 'Curriculum for Wales' policy is being implemented and, as a result, General Certificates in Secondary Education (GCSEs) for 14 to 16-year-old learners are being reformed. These new GCSEs will be more flexible, and will incorporate increased use of technology in assessment and more teacher assessment. The reform has opened a policy window in which changes to the standard setting approach are under debate. Ensuring a well-informed discussion about the advantages and disadvantages of different standard setting approaches is important and challenging. For example, across the United Kingdom many people believe that GCSEs are graded using norm-referencing, where in fact attainment-referencing approaches are used. This presentation will outline a qualitative study that explored Welsh teachers' and policymakers' understanding of how different approaches to standard setting for GCSEs would influence both the operation of the qualifications and their perception of them. We undertook semi-structured interviews with 27 participants, a mix of assessment system insiders and teaching professionals. The presentation unpacks our findings and provides suggestions for how to better communicate standards to stakeholders.

14:45 **Moderation - exploration of methodology for setting tolerances for general qualifications in England.**
Blake Ashworth

Abstract: For non-examined assessments in England, which are marked internally within centres, quality-assurance procedures like moderation must be carried out by awarding organisations which scrutinises raw marks awarded by centres and thus possible mark adjustments. Typically, tolerance levels of difference between a centre's mark and a moderator's mark on percentage scales are set by awarding organisations and are used for deciding if mark adjustments and additional script sampling is necessary. Regulatory bodies propose a maximum tolerance limit, but nowadays the typical level of tolerance is 6%. Tolerance levels set too wide risks overlooking poor marking in centres, but setting too narrow increases the burden on extra work for moderators and acceptable marks being adjusted. Our paper investigates whether there are any significant impacts on component and qualification outcomes by relaxing the tolerance levels. Our predictive model analysis and statistical testing from minimal tolerance increase shows minimal statistical impact of higher-grade outcomes and distributions, as well as significant increases in centres requiring no raw mark adjustments and reduction in scripts required for sampling. Therefore, some flexibility to tolerances can minimise the burden of moderation but not significantly impact overall outcomes.

Process Data II
Chair: Irene Custodio
Room: Christian Barnard (n=200)

13:45 **What can process data can tell us about students' persistence? Evidence from the e-TIMSS 2019 assessment**
Elena Papanastasiou, Evi Konstantinidou, Katerina Gkolia

Abstract: The digital transition in large-scale assessments generated a plethora of process data that can be utilized to provide additional information regarding examinees. Through data obtained from e-TIMSS (Trends in International Mathematics and Science Study) 2019 from 4th graders across 35 education systems in mathematics, our findings indicate significant variations on the time spent on multiple-choice items across different countries. However, emphasis will be placed on two countries, Germany and Denmark who have large variations in their overtime success and overtime failure variables, although their overall achievement in mathematics did not differ significantly from each other.

14:15 **From assessment of learning outcomes to assessment and support of learning processes: The role of process data in assessing and enhancing self-regulated learning**
Suijing Yang, Fabienne van der Kleij

Abstract: Traditional assessment design provides information about what students have achieved at a certain point in time but offers limited insights into how students learn over time. It poses challenges for educators to identify difficulties in learning processes and decide when and how to support them. With technological advancements, learning tools and platforms collect rich process data about students' learning. This data, when harnessed for assessment, holds the promise of enhancing our understanding of student learning and providing effective feedback. Yet, the potential and challenges associated with this innovative approach have not been fully explored. This conceptual paper focuses on the role of process data in assessing self-regulated learning and supporting students' self-regulated learning during and after assessments. We introduce a conceptual framework that outlines the potentials and challenges of leveraging process data. The framework is elaborated through three dimensions: (1) process data for assessing multi-faceted regulated learning, (2) time to leverage process data, and (3) signification of process data across different stakeholders. This framework advances current research by elucidating novel approaches for triangulating multimodal data in the assessment of learning processes. It has practical implications by providing evidence to guide the use of process data for students, teachers, and leaders.

International Assessments II
Chair: Therese Hopfenbeck
Room: Aphrodite B (n=50)

13:45 **The impact of open book exams on high school teaching and exam preparation after one year**
Rebecca Chivers, Vanessa Scherman, Rebecca Hamer

Abstract: Open book exams (OBEs) are expected to encourage a greater focus on higher order thinking in assessment performance. However, academic literature comprises mostly small-scale studies presenting mixed results and little evidence relevant to implementation in an international multi-lingual high school context. In 2022 an international awarding body recruited some 290 of their high schools worldwide to participate in multi-year quasi experimental study. This study compares the impact on learning, teaching, exam preparation and performance using multi-wave surveys and student grades. This paper will present the impact of three different types of OBE on the exam cohort that received one year of embedded teaching for OBE or traditional closed book exams (CBEs). At the start of the study, the students and teachers in OBE and CBE schools were similar in their level of wellbeing, growth mindset and learning and teaching experiences. Data collection is being completed at time of writing. Initial analysis of 10% of the data (50 teachers and 900 students) shows very similar patterns for CBE and OBE in student learning experience or strategies, teacher teaching and exam preparation practices and performance. Results of the full analysis of the impact after one year will be presented at the conference.

14:15 **PISA 2025 Foreign Language Assessment: The framework and science behind the test**
Angeliki Salamoura, Catalina Covacevich, Martin Robinson

Abstract: Proficiency in more than one language is a key asset for study, employability and cross-cultural communication in today's interconnected world. As a result, educational authorities worldwide are investing significant resources in foreign language teaching and learning (OECD 2020). But are their students achieving the expected language learning goals? The PISA 2025 Foreign Language Assessment (FLA) and its accompanying Framework were designed to provide evidence that will answer this fundamental question. The PISA FLA Framework defines foreign language proficiency and use; skills and competencies necessary to use a language; and the cognitive, social and cultural factors which influence successful language learning. It also describes how these competencies, skills and factors will be assessed and reported to provide the required evidence that will guide education policy decisions and ultimately the improvement of education for language students. We will discuss the design of the FLA, its Framework and the challenges inherent in the development of this large scale, international assessment and survey. In doing so, we will address the following questions: How will PISA FLA assess and report language learning outcomes? How will this information help guide education policy, contribute to JEDI (Justice, Equity, Diversity and Inclusion) and improve future outcomes?

15:15 - 15:45 Coffee Break

Foyer outside Akamas Room

Opportunity to visit SIG Banners

15:45 - 17:15 Ignite & Symposium Session

Ignite Session
Chair: Stuart Shaw
Room: Akamas A & B (n=550)

National Monitoring Tests in International Contexts
Anna Greene

Abstract: Cambridge University Press and Assessment has increasingly been involved in supporting Ministries of Education around the world in meeting their education reform agendas in relation to designing their national monitoring tests. We have seen an increase in international interest in relation to their existence, what they can be used for and how they can be designed, developed and administered. This presentation will draw on Cambridge's experience working with Ministries of Education on these tests. Areas of focus will be: • Trends in the drivers for having national monitoring tests • What the results being used for and the extent to which the validity of the tests can be at risk through multipurpose design • How advances in technology help and hinder the design and administration of the tests • The challenges involved in designing and administering the tests as well as interpreting and reporting results.

SARI: A New System for Automatic Reviewing of Multiple-choice Items
Séverin Lions, *Pablo Dartnell*, Abelino Jiménez, Matías Altamirano, Danner Schlotterbeck, Diego Reyes, Christian Collado, Laura Leal

Abstract: In this Ignite presentation, we will present SARI (System for Automatic Reviewing of Items), an intelligent system that automatically (and thus almost immediately) detects more than twenty item-writing flaws commonly reported in the literature. As far as we know, SARI is the most advanced system to complete this task for educational testing. SARI detects flaws for each item, highlights which text part is involved, and proposes a general solution to eliminate the flaws. All these tasks are done in a secured virtual environment. Flaw detection is based on recognizing elementary editing features, such as the presence of certain characters or words, and on the semantic and syntactic analysis of response options using automatic language processing techniques. SARI makes it possible to improve multiple-choice items and gain valuable time during the item reviewing process. It is thus potentially helpful for both teachers and standardized test developers.

Generating personalised feedback through low-stakes formative assessment
Maria Pereira, Manuel Gomes, Ana Monteiro

Abstract: The Institute of Educational Assessment (IAVE) is responsible, in Portugal, for external student assessment, both national (low-stakes and high-stakes testing) and international. Low-stakes tests generate performance result reports of students in different subjects at national, regional, school and class levels. They also provide individual reports for students, identifying strengths and areas for improvement. Coding procedures, which match students' responses to different performance levels, allow for descriptive and detailed feedback to each student, with recommendations for improvement in specific areas. Considering the potential of these reports for understanding students' knowledge and skills, IAVE has been working in partnership with schools (headteachers, teachers, students and parents) to optimize the analysis and use of these reports. Furthermore, IAVE has developed a Digital GPS (Guide of Practices and Suggestions) to assist all stakeholders in making the most of this information for learning improvement. Transforming individual standardised reports in e-format will be the next step.

Closing in on motivating computerized assessment: depicting the contours of the next generation of adaptive reading tests
Per Henning Uppstad, Bente Rigmor Walgermo

Abstract: Back in 1988 Bunderson et al. predicted four generations of computerized testing, derived from the technological foresights of that time. The foreseen generations included 1) the transfer of paper-test to computer, 2) computerized adaptive tests, 3) continuous measurement and 4) intelligent measurement. Since then, however, the view of the student's role in assessment and learning has changed radically, representing societal changes that challenges elements of the proposed generation 3 and 4 making these less plausible as precise descriptions of what lies ahead. In the present study, we discuss selected initiatives for digital assessment development as a case for the next generation of computerized adaptive testing. As a starting point, we first present the status and challenges of computerized assessment of reading in the Nordic countries relative to the Bunderson et al's foresight. Second, we predict the direction of the next generation of adaptive computerized assessment of reading.

Symposium: Elevating Student Agency in Assessment and Feedback in the AI Era

Chair: Fabienne van der Kleij - Discussant: Therese Hopfenbeck
Room: Akamas C (n=200)

15:45 **Use of Technology, Artificial Intelligence, and Process Data to Unlock Student-centred Assessment Feedback Practices at Scale**
Fabienne van der Kleij, A Therese N Hopfenbeck

Abstract: Research on assessment to inform and enhance learning has long recognised the power of feedback. Feedback is widely promoted as a low-cost, high impact strategy in all learning areas, for which there is strong evidence. Feedback may come from teachers, peers, learners themselves, or technology. Advances in technology have enabled facilitation of timely and individualised feedback processes at scale. Nevertheless, the intended impact of feedback is often not realised in practice. Many technology-assisted learning environments overlook contemporary feedback research, which emphasises the critical role of students as active agents in assessment and feedback processes. Research has yet to shed light on student processes of feedback use, a critical gap which can be addressed by technology and AI, and a focus on process data. This paper examines technological advancements in light of an evidence-based articulation of four categories of student role in feedback. These categories range from no student role to substantial student role. It highlights opportunities and pitfalls in utilising technology to advance research and practice in a student-centred feedback perspective. The paper highlights the need for interdisciplinary efforts in enabling effective feedback practices at scale, as well as attending to students' internal and external feedback processes through self-regulated learning.

16:15 **Oral Assessment and Student Agency in the Dialogic Space: How can AI Enhance Validity?**
Ayesha Ahmed, Chris Deneen

Abstract: Oral assessments are becoming popular in response to AI threats to authentication of written work, but beyond this they can offer unique opportunities for enhancing validity through dialogue between student and assessor. Establishing validity evidence traditionally involves a series of judgements that frame students' experience of assessment, rather than allowing them an active role in limiting threats to validity. If oral assessment is truly interactive, and exploratory talk is achieved during the process, then the act of assessment takes place in the dialogic space between student and assessor. It is here in which the process becomes collaborative. Misunderstandings can be mediated through feedback, as meanings are co-constructed, allowing construct relevant variance to be limited during the assessment act, and offering the potential for more authentic construct representation. The dialogue provides a window into students' thought processes, emotional regulation, and strategy use – key sources of validity evidence. There are major challenges of reliability, scalability, evidence capture, and oracy competence. The paper will highlight the potential for AI to address these challenges and enhance dialogic assessments through features such as automated scoring, feedback, and adapting the discourse.

16:45 **Exploring the AI-Education Nexus: A Scoping Review**
Anastasiya Lipnevich, A Therese N Hopfenbeck, Christopher DeLuca, Carmen Florentin

Abstract: With artificial intelligence (AI) rapidly permeating educational technology, a comprehensive understanding of existing research on AI's impacts on learning is crucial. This scoping review aims to map the current landscape of empirical studies examining links between AI implementation and educational outcomes across K-12 and higher education contexts. Following PRISMA guidelines, we conducted a systematic search of eight databases to identify relevant peer-reviewed studies on the use of AI and educational outcomes. The final set of studies was coded for outcomes assessed, AI technologies utilised, research methods, participant samples, and core findings. Analysis reveals AI is being applied to a diverse array of topics including intelligent tutoring, personalized learning, automated assessment, and more. However, research quality is uneven, with limited generalizability. Analysis reveals that studies tend to concentrate on short-term outcomes such as student perceptions, engagement, and performance metrics rather than longitudinal impacts on deeper learning or long-term educational trajectories. Our comprehensive scoping review highlights areas where high-quality, ecologically valid research is needed to optimally harness AI's potential in education. Findings can guide institutions, educators, and developers toward more rigorous and pragmatic inquiries that drive effective AI implementation. Strengthening the evidence base is imperative as AI's educational role rapidly evolves.

Symposium: Human connections for assessment in a technological age?

Chair: Lesley Wiseman - Discussant: Isabel Nisbet

Room: Christian Barnard (n=200)

15:45 **Developing multimodal assessment practices in technology-rich classrooms**
Henning Fjørtoft, Øystein Gilje

Abstract: Teaching and learning are increasingly taking place across various digital devices, with varying purposes, and students are using a range of semiotic resources (Kang, 2022) in their digital composing. Such rapid changes have impacted several aspects of human experience in education, creating frictions for students and teachers in the wider context of educational governance, hardware availability, and business models for educational technology (Nichols & Johnston, 2020). Drawing on data from a partnership between schools, local school authorities and university researchers, this paper studies teachers' and researchers' collaborative efforts in developing multimodal digital classroom assessments (MDCAs) for technology-rich classrooms (Author 2, 2020). We analyze frictions in developing MDCAs across four dimensions: (1) (lack of) textbooks, (2) classroom assessments, (3) curriculum areas, and (4) policy implementation, identifying tensions between the national framework for working with assessment and the growing number of new ways of demonstrating competence across a wide range of genres in technology-rich classrooms. These frictions impact issues such as the validity of classroom assessment, teacher professional development, and students' experiences and identity formation in educational contexts. We discuss how using digital technologies for assessment purposes affect student and teacher learning experiences.

16:15 **Preserving what is best in the role of the teacher**
Andrew Watts

Abstract: o Preserving what is best in the role of the teacher A teacher's role is fundamental to human life. The teacher could be anyone with some knowledge, perspective or skill which a learner wishes to take on. Where ever it takes place the task of teaching includes assessment. Especially in an era in which formative assessment has been advocated, the nature of the teacher–learner relationship is paramount. This talk dwells on aspects of the teacher–learner relationship. Biesta in a recent talk* has reiterated his thinking about what makes that relationship special. He notes that a fundamental action of the teacher is pointing - "Look there". But Implied by that gesture is the relationship "You look there". A machine could not replace that relationship. Learning is all about taking on something, making it part of yourself. It isn't just knowledge: there is also motivation, commitment and action. The best teachers arouse those in learners. Biesta not only identifies 'Socialisation' as a key part of education, but also what he terms "subjectivisation", the development of autonomous and independent individuals. *"Taking the angle of the teacher", talk for the Scottish General Teaching Council, webcast 9th February 2024.

16:45 **An ecosystemic research methodology: how to build an assessment culture which fosters creativity and empowerment?**
Nathalie Younès

Abstract: Whether we are talking about the evaluation of individuals, programmes, or institutions in Higher Education (HE), the most common practices are based on quantitative standardisation. This emphasis on performance, impact and ranking has been reinforced by both neoliberal policies and the development of digital technology. Those quantifying produces an alienating assessment, which is destructive of the creative potential of individuals and groups How could we escape from the culture of separation and the primacy of quantitative models by taking into account the perspectives of stakeholders (individuality of learners and teachers) and the various dimensions of the milieu? We will address this question by exploring the growth of hybrid education and the use of online learning in HE from the HyPES European research project whose main objective is to develop a typology of hybrid training environments and self-positioning tool. This development is done by integrating the points of view of both teachers and students, the dimension of learning assessment and the consideration of the institutional contexts. The paper could focus on the project's objectives and methodology in order to present the process of building an ecosystemic research methodology that places environmental dynamics and stakeholder perspectives at the heart of the investigation.

19:30 - 23:00 Conference Dinner

Location: Thalassa Hotel

Saturday, 09 Nov

9:00 - 9:45 Keynote Speech
Chair: Therese Hopfenbeck
Room: Christian Barnard Room

Prof. Chris DeLuca, Faculty of Education, Queen's University, Canada

9:45 - 11:15 Open Paper Session VI

Artificial Intelligence V
Chair: Sarah Hughes
Room: Hermes (n=30)

9:45 **Marking AI generated student work – how good is it and can humans tell?**
Rebecca Hamer

Abstract: When in November 2022 the ChatGPT interface for powerful Generative Artificial Intelligence (GenAI) was released to the public, its potential impact on the validity and authenticity of assessment became an immediate concern. A flurry of small-scale experiments posted alarming results that fed the growing frenzy. Early in 2023, an international awarding body commissioned the creation of 150 samples in different languages. Using as much of the ChatGPT generated output as possible, participants created 150 credible samples of two text-heavy high school level summative assessment tasks in different languages. The GenAI samples were then mixed with a matching sample of authentic student work and triple marked. Marking examiners were told that some of the scripts had been created using GenAI and they were asked to evaluate each script on credibility to be authentic, the quality of its referencing, and the likelihood and extent of GenAI output used in the script. This paper will present further results on the process and strategies used, evidence of language bias and features that human markers in this study identified which helped them differentiate between human and GenAI generated samples created with the use of ChatGPT3.5 and 4 with a high level of accuracy.

10:15 **Personalized Adaptive, Dynamic and Formative Assessment in Methods and Statistics Education**
Wilco Emons

Abstract: This presentation addresses an educational innovation project aimed at improving statistics education in higher education through the development of a personalized dynamic formative adaptive assessment tool. An important feature of the tool is the possibility for students to get tasks within a substantive context of their own choice. For example, the student may indicate that he/she wants to practice statistics within the context of child development. The tool then uses artificial intelligence (AI) and large language models to generate questions within this specified context including context-specific data examples. This tool should enable students to engage extensively with statistical concepts in a manner tailored to their individual learning requirements and aligned with their specific substantive interests, leading to deeper learning and a more effective transfer of the acquired statistical knowledge and skills to real-world challenges. While this tool primarily focuses on enhancing statistics education within higher education, the underlying ideas and principles discussed in the presentation extend this domain. Moreover, with this presentation we hope to give an inspiring example of how technological innovations may improve learning by combining AI with enhanced assessment methodologies such as personalized contextual assessment, innovative adaptive testing, and dynamic automated item generation.

National Tests & Examinations V
Chair: Nicky Rushton
Room: Leda (n=60)

9:45 **The new attainment test regime for last year primary education and the nationwide grading standard in the Netherlands.**
Stefan Jansen, *Natacha Borgers*

Abstract: In the Netherlands, secondary education consists of six cognitive levels. Primary school teachers initially provide a placement advice for either one of these levels. A second opinion is provided in the form of a mandatory primary school leavers attainment test. This attainment test is embedded in a system involving six test providers. All offered attainment tests should yield a minimum quality and provide comparable and valid results. Therefore, all tests undergo rigorous evaluation based on an assessment framework to ensure they meet minimum standards and apply one nationwide grading standard. In 2024 the first tests were administered under the new governance. Our paper will offer insights into the inaugural year of this new test regime and the nationwide grading standard in particular. We provide an overview of the framework and procedures used, and share our experiences in striving for comparability among the six tests. By comparing equating using IRT scaling and an equating method using populations (as a validation tool) we conclude that there are significant discrepancies between the two methods. We will discuss the implications of our findings for the future and possibilities of using multiple tests to come to comparable and valid results.

10:15 **School leaders' experiences from supporting primary school teachers' use of a national level, digital mapping test for numeracy**
Guri A. Nortvedt, Henrik Hung Haram, Oksana Kovpanets, Andreas Pettersen, Eren Sübül

Abstract: At the primary level, Norwegian schools use national level mapping tests to identify students in grades 1 and 3 at risk of falling behind in numeracy. While class teachers are responsible for administering and following up the assessments, school leaders are supposed to supervise and support their work. Traditionally, these assessments have been paper-based, and students have had teacher support when completing the assessments. Consequently, teachers had first-hand experiences that might help with analysing the assessment data. Starting from 2022, the tests were digitalised, changing how teachers might work with test administration and follow up activities. The aim of this paper is to discuss the experiences of school leaders (N=5) in supporting teachers in their work with the mapping tests. The presentation will address how the school leaders interact with teachers when they prepare for test administration, the extent to which and how school leaders engage with the teachers and assessment data and how school leaders understand their role. Understanding how schools move from paper-based to digital assessment, identifying and addressing possible issues and changes to roles and responsibilities are important to ensure effective digital assessment practices as we continue to move deeper into the digital domain.

10:45 **Linking Norwegian national tests with concurrent calibration using DIF analysis**
Ga Young Yoon, Anja Aigeltinger, Maoxin ZHANG

Abstract: The Norwegian National tests are conducted for all fifth-, eighth- and ninth-grade students every year. National tests measures students' basic skills in reading, numeracy and English. Tests consists different sets that are either unique for that year, or with common anchor items same over the years to measure the change in national trend of students' basic skills. Our research aims to investigate the differences between two linking approaches—fixed item parameter calibration and concurrent calibration—in analyzing data from the Norwegian National Tests from 2014 to 2021. We also investigate the influence of Differential Item Functioning (DIF) among anchor items across each successive two-year interval. We use data from Norwegian National Tests, with approximately 50,000 5th-grade students participating. Our analysis adopts the Item Response Theory (IRT) framework and includes two procedures: DIF analysis and linking analysis. Our results showed that many anchor items had DIF between years, resulting in huge differences in overall group mean for each year. Different models produced similar results, except for the years 2019 to 2021. Linking ensures comparability across different test versions and years. This study contributes to the ongoing discussion on the best practices for linking and DIF analysis in large-scale assessments.

Process Data III
Chair: Michalis Michaelides
Room: Plato (n=20)

9:45 **One way or another: alternative approaches to standard setting**
Lauren Miller, Ana Ulicheva, Sumita Ishaque

Abstract: Standard setting activities traditionally require multiple raters to review exemplar test taker responses alongside a test form and manually record a judgement on a separate application. Standard setting, requiring expert judgement, places a significant cognitive load on raters while in addition they need to balance multiple administrative tasks before and during the activity. We tested a hypothesis that an alternative, more holistic approach to standard setting could alleviate administrative aspects of standard setting and facilitate rater judgements by allowing them to focus solely on test-taker performance and reach judgements, instead of switching between applications. An online platform was used to create and host our standard setting activity. All raters involved were very familiar with manual standard setting activities but had not previously used this type of holistic platform for such activities. As well as feedback gathered during training and standardisation, raters completed a post standard-setting survey to answer the research question: Does the new method facilitate rater decision-making (in their perception) and if so, in what way? Raters assessed the advantages and disadvantages of using this new method. We will present these alongside a discussion on how process data generated though this approach could inform future activities and improvements.

10:15 **Illuminating Self-Assessment Cognition via Joint Display Integration of Multimodal Data**
Nathan Rickey, Ernesto Panadero, Christopher DeLuca

Abstract: Despite being a global priority for education, supporting student self-assessment remains a challenge across education contexts because the cognitive processes of students engaged in self-assessment are not well understood. In this study, we leveraged a multimodal approach to examine university students' (n=25) cognitive processes and emotions while they used rubrics and exemplars to self-assess and revise their writing. We collected eye tracking data, electrodermal activity data, and screen recordings while participants self-assessed and optionally revised their own reflective essays using a computer, with on-screen access to a writing rubric, two contrasting exemplars, and the essay task instructions. Participants then engaged in a gaze-cued think aloud protocol, generating data on their cognitive processes and emotions throughout the self-assessment process which we analyzed via an inductive approach. Using joint display, we integrated gaze durations and mean peak electrodermal magnitudes across six self-assessment resources (rubric, two exemplars, essay, self-feedback, instructions) with themes from the think aloud data—i.e., cognitive processes and emotions—to illuminate distinct cognitive processes and emotions that emerged when participants were looking at each resource. Rubrics and exemplars provoked distinct types of comparison processes and emotions, while essays provoked participants to prioritize important revisions. Insights provide inroads for supporting student self-assessment.

Assessment of Practical Skills lll
Chair: Ayesha Ahmed
Room: Aphrodite A (n=50)

9:45 **Comparing OSCE Performance in Medical Students Trained Online Versus Face-to-Face During the COVID-19 Era**
Nicoletta Nicolaou, Panayiota Andreou, Maria Perdikogianni, Alexia Papageorgiou

Abstract: The national lockdowns during the COVID-19 pandemic shifted teaching of clinical and communication skills online. In this study we evaluate the impact of online Vs face-to-face teaching on undergraduate pre-clinical medical student performance in Objective Structured Clinical Examinations (OSCEs). The performance of two student cohorts was compared: one cohort with online teaching (n=84) and one with traditional teaching (n=124). Descriptive statistics (mean, standard deviation) and inferential statistics (independent samples t-test for performance comparison; $\chi 2$ test of independence for pass/fail rates; $\alpha=0.05$) were used. The online cohort significantly outperformed the traditional cohort, with mean score of 79.6% compared to 67.4%, and pass rates of 98.9% versus 96.0% respectively. There was no significant difference between the average global rating between the two cohorts. This disparity shows that, while online learning can effectively improve specific academic skills, it may not entirely duplicate the comprehensive clinical experience afforded by traditional means. However, online learning may cater more effectively to different learning styles, particularly in teaching communication skills. Our findings contribute to the ongoing debate on effectiveness of online teaching in undergraduate medical studies. Institutions should consider integrating online teaching elements with traditional methods for a more flexible and effective learning environment post-pandemic.

10:15 **Learning outcomes as the mechanism of personalisation in CASLO qualifications: where are the limits?**
Latoya Clarke, Milja Curcin, Asteria Brylka, Paul Newton

Abstract: In England, many regulated vocational qualifications adopt the CASLO approach to support personalisation through qualification design. In this approach, the same detailed learning outcomes and assessment criteria (defining the content and standards) can be acquired and assessed in different contexts/formats and can be exchangeable despite not being identical, supporting the needs of different students and employers. However, the academic literature questions the extent of personalisation in some CASLO qualifications and its implications for reliability and fairness, knowledge/skill transferability and currency. This presentation explores the mechanisms enabling personalisation and the factors influencing its nature and extent. A document review was conducted exploring the characteristics of a sample of 6 CASLO qualifications, examining them in relation to their purposes, cohorts and other features that might promote or limit personalisation. Benefits and challenges of personalisation were also considered from the perspectives of organisations that design these qualifications and stakeholders, including teachers and students. This presentation will illustrate that the CASLO approach as a design template can be fluid and contextually responsive. It will also recognise the limits placed on the extent of qualification personalisation in different contexts given the need to establish a balance between educational value, reliability, validity and manageability.

10:45 **From Words to Wins: Refining our understanding of communication ability**
Sumita Ishaque, Ana Ulicheva, Rose Clesham

Abstract: Dialogue is one of the most common ways of communicating. It involves using language collaboratively to reach a common goal. Assessing dialogic ability should measure both the ability to use language, and to achieve an intended goal: a well-formed utterance is unsuccessful if it fails to achieve its purpose. Paradoxically, assessments focus primarily on characterising the quality of generated discourse but rarely on the effectiveness of interaction ability. This disregard for communication success is understandable: the concept is vague and subjective, not entirely clear in meaning or how to measure it. The main research question was: To what extent do common language tasks reflect the ability to communicate successfully? L1 and L2 speakers of English completed two key psycholinguistic tasks. The first was a referential communication task that involved describing abstract figures, called ‘tangrams’. In the second task, participants provided definitions for abstract and concrete terms. A separate group of ‘listeners’ were presented these definitions asynchronously to derive measures of communication success. Correlations between these measures and participants’ performance on a high-stakes language test were computed. We will discuss strengths and weaknesses of our approach and share our thoughts on the applicability of these findings to language testing constructs.

E-Assessment VI
Chair: Ben Stafford
Room: Christian Barnard (n=200)

9:45 **Designing a digital numeracy assessment for the 21st century**
Jeanne Marie Ryan, Hannah Rowe

Abstract: How would you define numeracy if you wanted to assess the skills that young people need to be successful in everyday life? What types of content would you want to cover, and what types of questions would you want to ask? These are the areas AQA and AlphaPlus have been investigating as we design a new digital assessment that will be used in order to support secondary students to master the numerical skills necessary for the modern world. This presentation will explore how we drew from two important sources to define constructs and refine assessment domains: from the academic literature on numeracy assessment and from the knowledge of individuals with decades of experience working in maths, in assessment and in areas related to numeracy in the professional world. We will also discuss some initial feedback received from students about their experience taking part in early trials of these digital items. Our aims are to share the experience of building a new digital on-demand assessment from the ground up and to encourage further conversation about how assessment creation functions as a community of practice.

10:15 **Accessibility and the use of diagrams in onscreen mathematics and science assessments for young learners**
Brooke Wyatt, Rebecca Conway

Abstract: Accessibility and inclusion are central considerations in the design of any assessment to enable equity of access for learners and to reduce the potential for construct-irrelevant variance. This proposal considers accessibility and inclusion in a novel assessment context – 'progress' tests in mathematics and science taken by a global cohort of young learners who speak English as a first or additional language (EAL). A previous research project on these tests, intended to identify language features that could present barriers to access for the EAL students, highlighted potential issues with the use of diagrams. This indicated a need to explore the use of diagrams and images to consider the specific accessibility challenges that they present and how they can be mitigated. The purpose of this presentation is to bring together the guidance for use of diagrams, colour and alternative text (alt text) from the accessibility experts in the field, while adding practical examples from our experience with assessments with young learners in an international context. The findings have been consolidated into guidance for our item writers and reviewers to support greater understanding of how to use diagrams and images in assessment tasks and their impact on onscreen accessibility.

Psychometrics & Test Develpoment III
Chair: Angela Verschoor
Room: Athena (n=60)

9:45 **Investigating the Comparability of Scenario-Based Equivalent Forms using Process Data: The Case of Digital Literacy Assessment**
Daria Gracheva, Ksenia Tarasova

Abstract: Modern education aims to develop and measure not only individual knowledge and skills, but also more complex constructs and literacies, such as digital literacy (DL). One of the most promising approaches to assessing complex constructs is virtual performance-based assessment (VPBA), where skills are assessed through students' observable behaviour in a test environment using simulations. However, there is a paucity of research on the development and comparability of equivalent scenario-based tasks as a type of VPBA. In the case of more interactive testing environments, analysis of the comparability of test results may include how the test taker solved the problem set for them, including the equivalent number of clicks made or time spent on each action. The general aim of our study is to develop equivalent forms of scenario-based tasks for the digital literacy assessment tool and to investigate comparability using performance and process data. Following the principles of automatic item generation studies, we aim to construct structurally and psychometrically isomorphic scenario-based tasks that are derived from the same template and have similar psychometric properties. The study is conducted using the example of the DL tool, which was created in the form of scenario-based tasks according to the Evidence-Centred Design approach.

10:15 **Causes of local item dependence in the SweSAT**
Per-Erik Lyrén, Inga Laukaityte, Christina Wikström

Abstract: Local item dependence, LID, is important to examine as it can affect item and test parameters as well as ability estimates. In this paper, we examine causes of LID among pairs of items in the SweSAT, a test for selection to higher education in Sweden. Because the SweSAT subtests differ in format, we also wanted to investigate whether LID is more prevalent in some subtests than others. We examined four test forms of the SweSAT, administered from spring 2022 to autumn 2023 (n ranged from 28,138 to 57,935). Furthermore, to detect LID we used linear partial correlations and Yen's Q3 index, with values of 0.1 and above indicating LID. In general, the number of item pairs in the test exhibiting LID was small. While content multidimensionality seems to be a cause of small LIDs, the most severe instances of LID were observed in two of the testlet-based subtests. The LID information is very useful as feedback to the item writers and test developers. First, it shows that they do a good job when it comes to avoiding redundant items. Second, by identifying causes of LID, remedial actions can be taken to minimize the risk of LID in future test forms.

10:45 **Combining multiple equating information sources**
Marieke Van Onna, Silvia Rietdijk

Abstract: The cut scores on the end-of-school exams of 2024 in the Netherlands will be equated to the cut scores of the exams in 2023. Multiple sources of equating information are collected and used: previous cut scores, expert judgements from Angoff standard settings, comparative judgments of the test difficulty by construction team members, comparative judgments of the test difficulty by teachers, score distributions, and population ability shift estimates based on IRT. Each source of information is translated into a cut score with a confidence interval. These source distributions are combined to one posterior distribution of the most suitable cut score. For the combination, a meta-analytical method is chosen over two other options (Bayesian and random effects). The pros and cons of the combination options are discussed.

Other IV
Chair: Ezekiel Sweiry
Room: Zeus (n=30)

9:45 **Introducing a New Self-Report Scale Format: Explicit Continuum Scale**
Inna Antipkina

Abstract: This study presents a methodology for developing a new format called 'explicit continuum scales' on the example of the Client orientation questionnaire. Elements of the Rasch-Guttman scenario scale methodology were used in its development. In three consequent studies different aspects of the scale functioning were investigated. The scale format demonstrated very high stability of dimentionality and item characteristics. Advantages and limitations of the format are described.

10:15 **The performance of transformer-based auto-markers on science content: a scoping review**
Frank Morley, Emma Walland, Carmen Vidal Rodeiro

Abstract: 'The transformer' is a model that allows computers to interpret written language by considering the context and relevance of words in sentences. This technology has led to the development of state-of-the-art natural language models such as GPT3.5, GPT4 and BERT. This presents opportunities for the auto-marking of science content, especially in formative, low-stakes assessments. This scoping review provides an overview of recent auto-markers (2017 onwards) for science questions, focused on assessing subject knowledge and understanding. We evaluate the performance of auto-markers with reference to both quantitative (e.g., Quadradic Weighted Kappa) and qualitative metrics (e.g., explainability and ethics). The review is underpinned by theories of what makes human markers effective and the implications of auto-marking systems seeking to imitate this. We followed a rigorous scoping review methodology, using Scopus as our search database. After applying our search terms, we screened 252 abstracts against our inclusion and exclusion criteria. This led to 20 studies being reviewed in depth. Our findings summarise recent developments in the auto-marking of science content, including opportunities and challenges.

Technical, Vocational & Applied Assessments
Chair: Hayley Dalton
Room: Aphrodite B (n=50)

9:45 **Assessing problem solving in Functional Skills Qualification Mathematics in England**
Diana Tonin, Becky Foster, Eve Taylor

Abstract: In England, Functional Skills Qualifications (FSQs) in mathematics are designed to provide learners with basic academic skills required in everyday life, enabling them to apply mathematical thinking to solve problems in familiar situations. The purpose was to deepen our understanding of how problem solving is assessed in this context, to inform Ofqual approach to regulating FSQ maths . FSQs in mathematics are strongly focused on problem solving. However, there is disagreement in the literature over how this is defined. To deepen our understanding of the assessment of problem solving in FSQ mathematics and therefore inform Ofqual's approach to regulation, we completed two studies looking at how problem-solving is assessed in FSQ maths. Using comparative judgment and a rating exercise completed by subject experts, we aim to identify the type of items that best elicit problem solving and their characteristics. In this presentation, the results of the research are presented and the features of items that elicit problem solving are discussed in the context of FSQs. The findings from this research provide Ofqual with a strengthen understanding of how problem solving is currently assessed in FSQ mathematics, allowing Ofqual to identify best practices and help ensure that the assessments are valid.

10:15 **Equating Functional Skills exams using Item Response Theory**
Zeeshan Rahman, Bas Hemker, Wobbe Zijlstra

Abstract: A pilot was undertaken to better understand how IRT could be used to equate examinations in a vocational setting. This involved creating three new versions of the mathematics Functional Skills exam with some common items between versions. Various test designs with common items were considered but the chosen design provided sufficient overlap between versions for IRT analysis but minimised item exposure. Items were carefully selected and evaluated to ensure their suitability for IRT equating. IRT pass marks for all three versions were derived and then compared with those set by the expert panel. Pass marks for two versions were at the same ability level but one was slightly different. However, the pass marks set by the expert panel were used to determine final learner results, as agreed with stakeholders upfront, given that the IRT approach was being piloted in this study. This paper presents the methodology and findings from this study and discusses potential benefits and challenges (with possible mitigations) of using the IRT approach for equating multiple exam versions. It also shows how technology can be used to enable efficient processing and analysis of data as well as transfer of skills and knowledge between organisations.

10:45 **Assessing behaviours in apprenticeship End-Point Assessments in England**
Fiona Leahy, Stephen Holmes, Nathan Pearson

Abstract: End-point assessment (EPA) is the final stage of an apprenticeship in England. It is an independent assessment to test that the apprentice has sufficiently demonstrated the knowledge, skills, and behaviours identified by employers as important for an occupation. Behaviours specifically are the mindsets, attitudes or approaches needed for competence in the role, for example, being “respectful of others” or “resilient under pressure”. However, behaviours tend to be more challenging to assess compared to knowledge and skills. An earlier strand of this research found that a range of different assessment methods were used to assess behaviours. The aim of the current strand was to understand how decisions around sufficient demonstration of behaviours are made in practice. Using an adapted think-aloud interview method, EPA assessors’ decision-making processes were observed whilst they watched video recordings of EPAs. A range of assessment methods were included in the recordings (including professional discussions, observations, and portfolios), in several different occupational areas. The results of a qualitative analysis of these interviews are presented. We discuss how behaviours are defined in the context of a need for reliable assessment, and the types of evidence assessors look for.

11:15 - 11:45 Coffee Break

Foyer outside Christian Barnard Room + outside terrace

Opportunity to visit SIG Banners

11:45 - 12:30 Keynote Speech
Chair: Elena Papanastasiou
Room: Christian Barnard

KTNRA Winner - Dr. Heather Kayton, University of Oxford, England

Evaluating the validity and comparability of PIRLS 2016 in South Africa

12:30 - 13:00 Closing Ceremony including Poster Award & Accreditation Awards 2025 Presentation
Chair: Elena Papanastasiou
Room: Christian Barnard

13:00 - 14:00 Lunch

Armonia Restaurant