51st Annual IAEA Conference

Toronto 2026

Trust, Transparency and Technology in Educational Assessment

Program



Host Organization

EQAO

Supporting Partner

Vretta

Sunday, September 27, 2026

10:00:00 - 17:30:00
Registration

11:00:00 - 12:00:00
Designing Game-based Formative Assessments of Mathematical Thinking
Location: Trinity Ballroom - Salon 3
Speakers: Dr. Robin Sharma
Game-based formative assessment can be an engaging, low-stakes, and dynamic method to evaluate mathematical learning processes. By embedding learning tasks directly into gameplay, educators can capture authentic mathematical thinking, problem-solving strategies, and conceptual understanding in real time through contextualized interaction. Moreover, game-based assessments can be particularly useful for mathematics, a discipline that is often associated with anxiety, stress, and pressure. This hands-on workshop will guide participants through the end-to-end creation of a customized digital minigame for their needs. Grounded in a research-based, three-prong educational game design framework, participants will select a math topic of their choice and work through a structured design process. Step by step, they will deconstruct learning objectives into progressive game levels, map learning and game mechanics, design authentic in-game puzzles, build immersive narratives, and integrate formative assessment mechanisms into their game design documents. Simple structured materials will be provided to guide the participants towards the creation of a holistic game design document. This workshop is for you if you have an idea of a math game but struggle to create a playable version for your students. Participants will create structured design materials (text to describe their game) that can be fed to an agentic AI tool to produce a customized, functional, web-based digital game tailored to their educational context. Targeted at math teachers, researchers, instructional designers, and assessment enthusiasts eager to modernize math evaluation, this session will equip participants with actionable design frameworks and prototyping skills. Pre-Requisites: Participants should bring a laptop to turn their designs into playable mini-games. They should have an idea for a math game that they want to create and have Claude Code or Antigravity installed on their computers.

11:00:00 - 12:00:00
Implementing Formative Assessments in Mathematics for Success in Post-Secondary Education
Location: Trinity Ballroom - Salon 5
Speakers: Dr. Ovidiu Bagdasar
Official statistics show that about half of the working-age population in England has numeracy levels lower than those expected from primary school children. This gap has an economic impact estimated at £20 billion in 2019, with the problem likely magnified recently by COVID-19 pandemic schooling disruptions, or by the over reliance on AI. In 2018, the University of Derby developed and implemented a formative assessment project designed on the principles of “assessment for learning”, that combines diagnostic assessment, interactive learning activities, and summative assessment to check learning, to help students to build confidence and competence with maths skills. The project has helped the University to improve Key Performance Indicators (KPIs) such as student performance, retention and employability, while the interactive dashboards help teachers to provide effective feedback to the whole cohort or to individual students, and university management to monitor engagement and progress. It also underpins the university-wide skills project Develop@Derby and enables training courses for regional SMEs and university partners within the UK and abroad. This collaboration has been chosen as Runner Up at the 2022 International eAA Conference under the “Best use of Formative Assessment” category, while the co-lead of the project, Professor Graham Orpwood received the Lifetime Achievement Award. The workshop aims to highlight lessons learned from this project, to explore avenues for improvement, and to identify potential adaptations which can support other learners.

11:00:00 - 12:00:00

When Feedback Becomes Affordable: Rethinking Classroom Writing in the Age of AI

Location: Bay

Speakers: Jean-Philippe Rivière

Writing occupies a central place in learning. Producing written work requires students to organize their ideas, mobilize knowledge, construct arguments, justify reasoning, and make their thinking explicit. Across many disciplines, writing is not merely an assessment activity, it is a learning activity in its own right.However, opportunities for extended writing remain relatively limited in everyday classroom practice. Longer or more open-ended written productions are often reserved for a few key moments during the school year, while more frequent assessments tend to rely on formats that are quicker to mark. This situation is not necessarily the result of pedagogical choices. It largely reflects a well-known constraint: the cost of providing meaningful feedback.Reading each piece of work carefully, annotating it, writing individualized comments, and assigning a grade requires a substantial investment of teacher time.This observation naturally leads to a simple question: How might teaching practices change if the cost of providing feedback on students' written work were significantly reduced, while teachers retained full pedagogical responsibility?Today, this question is becoming increasingly relevant with the emergence of AI-assisted feedback systems capable of analyzing written productions.This workshop therefore invites participants to explore a possible future in which AI contributes to the feedback process. No technical discussion is required. Instead, participants are encouraged to anticipate the pedagogical changes that such a development could make possible.

13:00:00 - 15:00:00

Critical Thinking and Creativity, Hiding in Plain Sight: AI-Supported Mapping of Complex Skills in Current Assessments

Location: Trinity Ballroom - Salon 5

Speakers: Kristin Lansing

In the age of AI, critical thinking and creativity have become essential capabilities for success. Many educational assessment and learning organizations cannot afford, or simply do not need, to buy or build entirely new assessments dedicated to these constructs. Opportunities for examinees to demonstrate complex thinking are already embedded in existing tests, tasks, and learning interactions; they are hiding in plain sight but rarely accounted for, evaluated, or leveraged systematically. This workshop introduces an AI-enabled, evidence-centered framework for making critical thinking and creativity visible within current educational assessment ecosystems and for strengthening these opportunities while preserving the primary construct being measured. We show how AI can support developers not only by generating tasks, but by analyzing prompts for reasoning, evaluation, synthesis, problem framing, perspective taking, and idea generation, and ensuring those demands remain construct-aligned rather than becoming unintended dimensions of measurement. Participants will learn a practical methodology for mapping where complex thinking is already elicited, judging the quality of those opportunities, and using AI to refine prompts and evidence models to produce richer, more authentic interactions, all without increasing testing time, adding new assessments, or overburdening faculty and staff.

13:00:00 - 15:00:00

Trustworthy Data Begins with Accessible Design: Inclusion Challenges in Digital Large-Scale Assessments

Location: Trinity Ballroom - Salon 3

Speakers: Elodie Vezon

As large-scale assessments increasingly move to digital environments, accessibility can no longer be treated as an additional accommodation layer. It must be embedded in assessment design, item development, platform features, administration protocols and data interpretation. This interactive session will explore how needs-based accessibility can strengthen both equity and measurement quality in technology-enabled assessment systems. The workshop will begin with three case-based stations. Participants will examine digital assessment situations involving learners with distinct access needs: visual access and navigation barriers, language and communication barriers, and neuro developmental or sensory barriers affecting task initiation,

attention and response production. Each case will invite participants to identify what the assessment is intended to measure, what barriers are introduced by the digital format, and which adaptations preserve the targeted construct. The second part of the session will connect these practical cases to broader principles for inclusive digital assessment. We will discuss how to distinguish access to information, mathematical or cognitive reasoning, and response production; how to document accessibility features at item and platform level; and how to ensure that accommodations do not alter the construct being measured. Particular attention will be paid to the consequences for large-scale assessment methodology: incomplete data, underestimated performance, missing responses, sample representation, comparability and validity.By the end of the session, participants will have identified key design principles for accessible digital assessments and reflected on how inclusive design can improve both fairness for learners and the quality of assessment data.

13:00:00 - 15:00:00
Designing Large-Scale Assessments to Inform Policy Reforms
Location: Bay
Speakers: Amy Dresher, Terran Brown
Large-scale assessments are designed to have little or no consequences for individual participants. Rather, samples of respondents participate in the assessment and results are reported at the group level. Oftentimes the cognitive assessment is accompanied by survey (or background) questionnaires that gather information about focal areas of interest. The design of the test can determine what results can be presented, what comparisons can be made, and what conclusions can be meaningfully drawn. This workshop will review crucial research-based design principles that can help researchers and practitioners design better large-scale assessments that policymakers and stakeholders trust. The workshop will start with an overview of large-scale assessments, followed by a review of design principles that promote equitable assessment and a presentation of case studies from several assessments. The workshop will conclude with a discussion of challenges and successes that participants have faced and may encounter as technology, AI and related factors continue to influence assessment. This workshop is designed to provide valuable insights for researchers, practitioners, and policymakers with interest in improving existing large-scale assessments or setting up new large-scale assessments in their countries from scratch. While some prior knowledge in large-scale assessment or survey methodology would be helpful, it is not required.

17:30:00 - 19:30:00
IAEA 2026 Welcome Reception

Monday, September 28, 2026

07:00:00 - 16:30:00
Registration

07:00:00 - 08:30:00
Breakfast

08:30:00 - 09:15:00
Opening Ceremony
Location: Grand Ballroom

09:15:00 - 10:00:00
Reimagining Assessment in the AI Era: From Assessment Triangle to Assessment Ecology
Location: Grand Ballroom
Speakers: Dr. Eunice Eunhee Jang
For over two decades, the assessment triangle (Pellegrino et al., 2001) has provided a foundational framework for understanding assessment through the interdependence of cognition, observation, and interpretation. Artificial intelligence now transforms all three simultaneously, changing the constructs we assess, the evidence we generate, and the inferences on which assessment decisions depend (Messick, 1989; Kane, 2013). The challenge is no longer how to incorporate AI into assessment, but how to rethink assessment itself. I argue that the assessment triangle remains foundational but must now be situated within an assessment ecology, a dynamic network in which learners, educators, AI systems, institutions, and policy continuously shape evidence, interpretation, and decision making. Because changes to one component reshape the others, designing trustworthy assessment requires systems thinking rather than isolated technological innovation. Drawing on a decade of BalanceAI and APLUS research across AI-supported classroom, language, and large-scale assessment, I propose an ecological framework articulated through three movements: toward equity, by shifting from ranking learners to profiling learning and from fairness to equity by design; toward transparency, by making learning processes and assessment systems—not only products and algorithms—visible; and toward trust, through principled AI integration and sustained human-AI partnership. Rather than treating AI as a threat or an add-on, this keynote argues that AI fundamentally reshapes the ecology of assessment. Ultimately, the question is not how powerful AI becomes, but how wisely we design assessment ecologies that generate richer evidence, more defensible interpretations, and more equitable educational decisions.

10:00:00 - 10:25:00
Coffee Break with Exhibitors

10:25:00 - 11:25:00
Global Perspectives on Assessment Modernization: Balancing Technological Innovation with Public Trust
Location: Grand Ballroom
Speakers: Anand Karat, Confidence Dikgole, Dan Koenig, Dr. Mary Pitoniak, Dr. Naveed Yousuf, Dr. Wayne Wesley, Pascal Couture
Purpose: Around the world, educational assessment bodies are navigating complex shifts to modernize their systems, driven by rapid technological advancements, changing socio-economic needs, and growing demands for systemic equity. This international panel brings together assessment leaders from Canada, South Africa, the Caribbean, Pakistan, and the United States to examine how different jurisdictions navigate the opportunities and challenges of assessment modernization while safeguarding public trust, data integrity, and validity.

Methodology: Chaired by Anand Karat (Vretta), this session features structured contributions from six panelists representing diverse regional contexts: Dan Koenig (EQAO, Canada), Confidence Dikgole (IEB,South Africa), Dr. Wayne Wesley (CXC, Caribbean), Dr. Naveed Yousuf (AKU-EB, Pakistan), Dr. Mary Pitoniak (Pitoniak Educational Measurement Consulting, USA), and Pascal Couture (Alberta Ministry of Education,Canada). The panel will explore regional drivers, technical logistics, change management, and psychometricintegrity during large-scale transitions. A moderated discussion and audience Q&A will follow.Key Findings/Insights: Panelists will share practical insights into balancing technological innovation with culturalresponsiveness, data security, and stakeholder trust. Attendees will gain a comparative understanding of howmulti-country examination boards, provincial agencies, and state testing authorities manage digital transitionswithout compromising equity or measurement standards. Key themes will include teacher engagement,continuous data flow, adaptive environments, and maintaining rigor during system-wide transformation.Implications for Educational Assessment: By providing a cross-jurisdictional lens on systemtransformation, this session highlights practical frameworks for international assessment leaders. It emphasizesthat successful assessment modernization extends beyond technological adoption to encompass changemanagement, institutional trust, and psychometric integrity.

11:30:00 - 12:00:00

Can LLMs Guess How Students Will Respond?

Location: Bay

Speakers: Christopher Ormerod

This session reviews several recent studies in which generative language models are used to simulate studentresponses in order to estimate item parameters before field testing. We explain the training mechanisms thesemethods share, parameter-efficient fine-tuning and ability-conditioned response simulation, and contrast twoconcrete approaches: one that generates open-ended written responses and one that reconstructs the optiondistribution of multiple-choice items. We summarize their accuracy against item response theory benchmarks,discuss validation evidence on how closely simulated responses resemble authentic student writing, and outline a Markov-chain extension to multiple-select items. Throughout, the central question is whether languagemodels can reproduce not only correct answers but the characteristic mistakes that students make.

11:30:00 - 12:00:00

Expert Judgment vs. Empirical Evidence in Item Difficulty Estimation: A Case Study of Aga Khan University Examination Board

Location: Trinity Ballroom - Salon 1

Speakers: Dr. Naveed Yousuf, Munira Muhammad Rangwala

High-stakes examination systems rely on the estimation of item difficulty to maintain consistency in assessmentstandards and support the overall quality of examinations over time. While many international testing programsemploy statistical pre-testing to calibrate item difficulty, such approaches are often constrained in examinationcontexts where confidentiality, security and resource limitations restrict piloting of items. In such contexts,expert predictive judgment is commonly used to estimate item difficulty during item development and standard-setting processes. This study examines the relationship between expert-predicted and empirically derived item difficulty in high-stakes examinations conducted by the Aga Khan University Examination Board (AKU-EB) in Pakistan. The study focuses on science subjects examinations in grades IX and X. A content expert at the examination board and two external schoolteachers with three to five years of teaching experience of the same subject and grade levels were engaged for predicting item difficulty. The study evaluates the effectiveness of predictive judgement by experts against empirically derived item difficulty using Spearman's correlations. The study also examines the correlation in predictive accuracy between exam board content experts and schoolteachers. The correlation between the predictive item difficulty using content experts and empirical item difficulty was moderate and statistically significant (? = .401, $p < .01$). The findings revealed that the correlation between exam board content experts and schoolteachers are ? = .179 indicating relatively weaker agreement across raters. Further analysis revealed that the judgement of the schoolteachers was more accurate towards the empirical item difficulty which reflects the knowledge of the context. The findings suggest that integrating a more rigorous approach towards achieving a difficulty estimate is capacity development by

sharing data regarding difficulty trends before the exercise and going through reinforcement on the factors that contribute towards item difficulty would help enhance the accuracy of item prediction in high-stakes

11:30:00 - 12:00:00 | **Open and Transparent Scoring with measr: An R Package for Diagnostic Classification Models**
Location: King
Speakers: W. Jake Thompson
Diagnostic classification models (DCMs) offer a powerful framework for moving beyond single total scores to provide actionable, attribute-level feedback about student learning. Yet the complexity of these models has historically made them a black box for practitioners and stakeholders: results are reported without clear communication of model evidence, scoring uncertainty, or the reliability of classifications. This opacity undermines the very trust that diagnostic assessments are meant to build.This presentation introduces measr, an open-source R package designed to make DCM-based scoring transparent, reproducible, and accessible. Built on Stan, measr provides a principled Bayesian estimation framework for fitting a wide range of DCMs, surfacing the evidence behind every scoring decision. Three features of measr are highlighted as pillars of transparent assessment practice:First, measr supports probability-based score reporting in addition to binary mastery/non-mastery classifications with posterior mastery probabilities, giving score users a direct, interpretable measure of classification uncertainty. Rather than a student simply "passing" a threshold, teachers can see that a student has, for example, an 87% probability of mastering a given attribute. This probability allows users to see who a student is to a given mastery threshold and understand the confidence in the classification. Second, measr provides a comprehensive suite of reliability indices, including classification consistency and accuracy, allowing agencies to quantify and communicate the dependability of score-based decisions. These indices are first-class outputs, not afterthoughts, making it straightforward to include them in technical reports and score interpretation guides. Third, because measr is fully open-source and integrates with standard R workflows, every step of the scoring pipeline from model specification to result extraction is scriptable, version-controllable, and auditable. This reproducibility enables agencies to demonstrate, not just assert, that their scoring procedures are sound. Attendees will leave with an understanding of how measr can be adopted within existing assessment workflows to strengthen stakeholder trust through transparency, and with practical guidance on incorporating probability-based reporting and reliability evidence into operational practice.

11:30:00 - 12:00:00
When Theory and Practice Align: Designing Effective Assessment for Early Numeracy
Location: Trinity Ballroom - Salon 4
Speakers: Dr. Heather Douglas
Transparency in assessment requires strong links between the researchers who develop assessments and the educators who will use them. An effective assessment tool can help educators to both identify students who need additional support to succeed, and to monitor their students' progress in response to intervention. The assessment should be based in theory (construct validity) and make sense to the educators who will use it (face validity). When used within a system-wide universal screening plan, effective assessment can build equity when all students are tested and those who are identified receive targeted support. In this session, I will discuss the development and implementation of the Early Math Assessment @School (EMA@School) in Canada.In 2021, researchers from Carleton University developed the EMA@School to measure the foundational numeracy skills that, according to the cognitive science of math development, are required for success in math. This screening tool has subsequently been used by over 200 000 students in kindergarten to Grade 4. In this session I will describe initiatives to build educator knowledge and also share examples of how educators have used the tool to test the efficacy of system-wide numeracy interventions. The EMA@School is a reliable and valid assessment tool that can be used to identify students who need support in math and to monitor student progress. Transparency and open communication between researchers and educators have supported the validation of the tool and enabled educators to target and test their interventions. The result is improved student success. For example, one school district found a reduction from 72% to 37% of intervention students needing additional support after one cycle of targeted intervention. Implications Building assessment transparency through shared knowledge and strong research and education partnerships has been a key factor for implementing successful system-wide early numeracy screening and intervention.

11:30:00 - 12:00:00

Beyond the Data: Building Trust and Transparency Through Black Community Partnership in Provincial Assessment

Location: Trinity Ballroom - Salon 2

Speakers: Deborah Buchanan-Walford, Nadea Clarke-Bartley, Pershia Nelson, Tamara Phillips

Efforts toward educational equity in large-scale assessment require more than technical precision. They require relationships. This session is a storytelling-focused reflection on the ongoing journey of the Education Quality and Accountability Office (EQAO) in Ontario as it works alongside Black community partners to build trust, transparency, and shared understanding in provincial assessment. Rather than offering a finished model or claiming solutions, EQAO Equity Education Officers share the honest, unfinished story of developing community partnerships. Participants will hear how trust is built over time, where tensions and uncertainties remain, and what it means to stay engaged and “at the table” while learning alongside the community. The session offers candid reflections on opening institutional processes to community voice, with the understanding that missteps, discomfort, and course correction are part of meaningful change.

11:30:00 - 12:00:00

Examining Predictability in High-Stakes School-Leaving Examinations: Evidence from the South African National Senior Certificate

Location: Carlton

Speakers: Matsie Agnes Mohale

High-stakes examinations play a critical role in regulating access to further education and employment. In South Africa, the National Senior Certificate (NSC), administered primarily by the Department of Basic Education (DBE) and independent assessment bodies, serves as the principal school-leaving qualification and is central to national development objectives. Recent upward trends in learner performance in some NSC subjects have prompted concerns regarding the potential predictability of examination papers and its implications for assessment validity. This paper reports on a qualitative document analysis of DBE NSC examination question papers administered between 2022 and 2024. Using a purpose-designed conceptual framework grounded in content and construct validity, the study examined ten high-enrolment subjects to determine whether repeated patterns in topic selection, item types, skill demands, and contextual framing contributed to unacceptable levels of predictability. Data was generated through expert review of examination papers and was analysed thematically across six dimensions of assessment design. The findings indicate that, across most subjects, DBE examinations demonstrated acceptable levels of predictability aligned with curriculum prescripts, without evidence of systematic repetition that would enable accurate anticipation of examination content. However, two subjects, Mathematics (Papers 1 and 2) and English First Additional Language (Paper 3), were found to exhibit unacceptable levels of predictability. These findings suggest that predictability in the NSC is subject-specific rather than systemic. The paper concludes by discussing implications for examination design, quality assurance, and international debates on balancing transparency and variability in high-stakes assessments.

11:30:00 - 12:00:00

Data with Purpose: Building Trust Through Instruction-Centered Assessment Systems at the Peel District School Board

Location: Trinity Ballroom - Salon 5

Speakers: Denise Gilliland, Laura Richards

As large-scale assessment data continue to be used for accountability purposes, educators increasingly seek tools that translate data into meaningful instructional support. This session explores the Peel District School Board’s (PDSB) journey in designing and implementing its digital EQAO Math Success Hub, a platform developed to balance accountability demands with immediate, practical instructional support for educators. PDSB’s EQAO Success Hub originated from a desire to provide students with authentic, deliberate practice opportunities aligned to curriculum expectations while simultaneously supporting educators in making real-time, data-informed instructional decisions. Moving beyond traditional post-assessment reporting, the Success Hub was intentionally designed to deliver in-the-moment data that educators can use to adjust

instruction responsively throughout the school year. This shift reframes assessment as an ongoing process that supports learning, positioning year-end performance as the cumulative result of continuous, informed instructional action. Participants will be guided through the design process of the Success Hub, highlighting how assessment data are transformed into actionable insights. Interactive data reports identify potential student misconceptions and pair them with high-impact, ready-to-use instructional practices that can be implemented immediately in classrooms. These resources reduce time demands on educators while reinforcing effective math pedagogy. Beyond data reporting, the Success Hub supports the development of teacher content knowledge for teaching mathematics. Embedded supports, including ready-to-use lessons, strategies for strengthening mathematical vocabulary, and examples of best math practice, build educator confidence and capacity to respond effectively to student needs. Teachers report feeling supported, contributing to a culture of trust where assessment data is viewed as a tool for growth. Preliminary findings from the Peel District School Board's Research Department indicate a positive association between student participation in the Math Success Hub and achievement on Provincial EQAO assessments, with students who engaged in the platform demonstrating higher rates of success compared to those who did not. This session demonstrates how PDSB'sDigital EQAO Math Success Hub reimagines large scale assessment data as a formative, instructional resource, supporting responsive teaching, strengthening professional practice, building trust in data use, and improving student achievement.

12:00:00 - 13:00:00
Lunch

13:00:00 - 13:30:00 | **Closing the Gap Between Evidence and Action**
Location: Grand Ballroom
Speakers: Steve Brown
Drawing on his experience leading transformation across multiple industries and now as CEO of Nelson Education, Steve Brown explores why the next evolution of assessment depends on strengthening the connection between evidence, curriculum, and instruction. High-quality assessment data is only one part of the equation. Lasting impact comes from helping educators understand what the evidence means and providing practical guidance on what to do next. As standardized assessments and universal screeners continue to evolve, there is an opportunity to move beyond reporting results toward supporting better instructional decision-making. Organizations that combine assessment expertise, curriculum knowledge, learning science, and instructional design are uniquely positioned to help education systems translate evidence into action. This session explores how a more connected approach can help educators move from identifying learning needs to confidently responding to them, ultimately improving outcomes for every learner.

13:30:00 - 14:00:00
Securing the Chain: Data Privacy and Exam Integrity in Distributed Digital Assessment Environments
Location: Grand Ballroom
Speakers: Sanketa Anand
Privacy and Security in Technology-Enabled Assessment As assessment delivery decentralized - across test centers, remote proctoring, and pop-up venues - the attack surface for exam security threats grows. This presentation examines the end-to-end data protection architecture required for distributed assessment:encrypted content packaging, biometric identity verification, off line response handling, and secure synchronization. Using real deployment case data to illustrate where traditional security models break downand what governance frameworks address the gap.

14:10:00 - 14:55:00
From Cheat Codes to AGI: Rethinking Assessment Security for an AI-Forward World
Location: Trinity Ballroom - Salon 3
Speakers: Paul Muir, Tim Burnett
This session is a direct call to action. We intend to challenge the assessment community to fundamentally rethink how we design and secure assessments before the next wave of AI makes our current approaches obsolete. The presentation is led by two industry leaders:Tim Burnett is former Chair of the e-Assessment Association, Founder of the Test Community Network and a recognised AI in assessment expert working with awarding organisations

and assessment providers globally. Paul Muir is Chair of the Board of ATP, former Chair of the ATP Security Committee, and a founding member of ATP's AI Committee. Between them, they bring decades of experience spanning applied AI, assessment technology and test security, making this a uniquely authoritative pairing to lead this conversation. The threat landscape has changed. AI is no longer simply generating essays. It is powering a rapidly expanding ecosystem of wearable devices, purpose-built cheating applications and intelligent tools designed specifically to circumvent assessment security. The sector is responding to yesterday's threat whilst tomorrow's is already being built. The solution is not to patch existing frameworks but to redesign assessment for the world AI is creating. Drawing on a human-first, AI-forward philosophy, Tim and Paul will present a practical framework for building agent-ready assessment environments. These are systems that test what genuinely matters in a world where AI assistance is a permanent feature of professional life, and where validity, security and authenticity must be rethought accordingly. This matters because the gap between current practice and the capabilities of emerging AI is widening faster than the sector is moving, both in terms of regulation and detection. The trajectory towards artificial general intelligence is not a distant concern. It is the context in which we are already operating. The window to act thoughtfully is now. Delegates will leave with a clear-eyed view of the threat trajectory and a set of actionable principles for beginning to future-proof their assessment systems.

14:10:00 - 14:55:00

What If Public Institutions Stopped Reinventing the Wheel? Nine Years and Still Building a Trusted International Digital Assessment Library Together

Location: Trinity Ballroom - Salon 5

Speakers: Amina Afif, Thierry Rocher

In 2017, four colleagues from four countries sat together at an international conference and asked each other a simple question: how are you coping with the transition to digital assessment? What followed was not a grant application or a strategic plan. It was a conversation — honest, peer-to-peer, and revealing. Every system was solving the same problems, spending the same resources, and largely doing it alone. That conversation became the FLIP+ Library. Now in its ninth year, FLIP+ brings together 22 public institutional members across six active Working Groups, with a shared, non-commercial, open-source platform for trusted digital assessment content — built by public institutions, governed by public institutions, and designed to remain a trusted international public good long after any single grant cycle ends. This session offers an honest, practitioner-led account (based on the experience and created output) of what it truly takes to build the FLIP+ Library, a digital infrastructure built by educators, trusted by systems across borders in public education and shared for all. We will share what the community has produced: a working platform now open in Beta Access, real assessment content piloted in national systems across multiple countries, and a governance model rooted in reciprocity and shared ownership. We will also name what it costs: sustaining volunteer momentum over years, navigating institutional inertia in a landscape where AI moves faster than public systems can, and keeping a collaborative initiative alive when it is (by nature) never anyone's single institutional priority. This is not a polished success story. It is a live account of a community that started with trust, kept going through a pandemic, and is still building, because the problem it set out to solve has not gone away. Participants will leave with a clear picture of the FLIP+ model, its real challenges, its open questions, and a genuine invitation: if your institution is solving these problems alone, expensively, and from scratch, you do not have to.

14:10:00 - 14:55:00

Bridging Alternate Assessments with Enhanced Reporting for Inclusive Employment Pathways

Location: Carlton

Speakers: Dr. Meagan Karvonen, Dr. Melissa Gholson

Students with significant cognitive disabilities (SCD) globally face profound challenges transitioning to post-secondary life, primarily due to a critical disconnect between functional-focused academic instruction and the comprehensive preparation required for meaningful vocational engagement. This presentation introduces the Postsecondary Opportunities (PSO) project, designed to forge clear, actionable pathways to inclusive employment for SCD learners. This initiative directly addresses the IAEA Conference themes of trust,transparency, and technology, with an emphasis on transparency and a strong focus on equity and fairness in Educational Assessment. The PSO project began with a robust mixed-methods needs assessment. Surveys (N=342) and focus groups engaged special educators across the U.S., gathering in-depth insights into

current instructional practices, barriers to career readiness, and resource needs. Building upon these findings, the project developed an innovative "Bridge database." This tool was crafted by systematically linking U.S.-based Dynamic Learning Maps (DLM) alternate academic standards to specific occupational requirements and knowledge, skills, and abilities (KSAs) derived from the comprehensive U.S. O*NET database. The needs assessment revealed a pervasive instructional disconnect despite educators' strong motivation to foster career readiness, they are severely hampered by a lack of integrated, standards-aligned resources. Key barriers included insufficient funding, limited transportation, employer hesitancy, and a tendency to treat academic instruction and career preparation separately. This insight underscored the urgent need for a systematic tool that explicitly functionalizes the relationship between assessed academic knowledge and concrete career-ready abilities. The "Bridge" tool addresses this gap, transforming abstract academic standards into tangible vocational relevance. Implications for Educational Assessment: By integrating transparent occupational alignments directly into accessible individual student reports, this project promotes clear,accessible, and responsible communication of assessment results to learners, their families, and multidisciplinary support teams. Transforming traditional academic scores into actionable career-readiness insights directly supports collaborative, individualized transition employment goal setting. Ultimately, the Bridge champions greater impartiality and fairness in assessment by creating equitable pathways. This fosters trust among stakeholders by providing concrete support for student growth and development toward diverse,inclusive employment opportunities.

14:10:00 - 14:55:00

Educator Voice and Assessment Reform: A Latent Profile Approach to Differentiated Gen AI Guidance

Location: Bay

Speakers: Jessica Yarbro

Formative assessment (FA) is a cornerstone of effective teaching across ages and stages, but student over-reliance on Gen AI is increasingly viewed as a threat to learning. Further, educators vary in their sentiment and familiarity with Gen AI, creating uncertainty around best practices for Gen AI in classrooms and schools. The purpose of this research was to understand differing educator perceptions of the impact of Gen AI on FA and to identify promising approaches for preserving learning integrity in this era. To achieve this aim, we surveyed 1,000+ US and UK educators working in secondary schools or Higher Education (HE). Latent Profile Analysis was applied to identify educator profiles based on survey response patterns (specifically, responses around FA use, Gen AI sentiment, Gen AI familiarity, and perceived student GenAI familiarity). We also interviewed 15 global educational expert practitioners to gather recommendations for FA redesign; a mix of inductive and deductive content analysis was used to analyze these data. Across our sample,?the FA tasks educators reported using most often (e.g., essays, multiple-choice questions) were also those perceived as most vulnerable to Gen AI misuse. Sentiment, familiarity with GenAI, and changes in practice varied across our sample, and four distinct educator profiles emerged: "Staunch Traditionalists”- most negative and skeptical toward GenAI in the classroom “Protective Skeptics”- most protective of FA, but open to GenAI’s possibilities “Cautious Explorers” - neutral, but curious about the role of Gen AI in their FAs, and “Proactive Innovators” - positive toward Gen AI and most experienced with Gen AI integration in their FAs Expert practitioners highlighted different points along the AI integration journey, recommending practices for redesigning FA to preserve learning integrity, including incorporating collaboration, focusing on process over product, and diversifying assessment formats and modalities. Recommendations are organized by educator profile to provide practical starting points that meet educators at their current level of Gen AI familiarity. These recommendations can help educators minimize risks and leverage opportunities afforded by Gen AI, which will help to maintain learning integrity and trust between educators and students more effectively than banning,policing, or reverting completely to in-class assessment.

14:10:00 - 14:55:00

Adaptive Item Targeting, Student Engagement, and Trustworthy ScoreInterpretation in Digital Mathematics Assessment

Location: Trinity Ballroom - Salon 2

Speakers: Blair Lehman, Dr. Jennifer Hove, Hyunah Kim, Mitch Haslehurst, Zhimei Gu

Student engagement is essential to the trustworthiness of large-scale assessment results, yet it is often treated

as a background assumption rather than a design consideration. This paper examines how item targeting in multistage adaptive testing relates to students' willingness to engage with assessment tasks. Drawing on process data from Ontario's Grade 9 mathematics assessment, we explore whether students are more likely to respond effortfully when item difficulty is appropriately matched to their proficiency. The study uses response-time data from a population-level digital assessment delivered through a multistage adaptive design. Engagement is identified through rapid-guessing indicators, and item targeting is examined through the distance between item difficulty and student ability. Rather than focusing on technical model details, the presentation highlights what these process data reveal about how students experience adaptive assessments and where disengagement is most likely to occur. Findings show a clear pattern: students are generally more likely to stay engaged when items are challenging but still attainable, and disengagement becomes more likely when items are substantially harder than their ability level. This pattern is especially pronounced for lower-achieving students, while higher-achieving students tend to remain engaged across a wider range of item difficulty. The results suggest that adaptive assessment does more than improve measurement efficiency; it can also shape the quality of student participation. The paper speaks directly to the conference theme of Trust,Transparency and Technology. From a trust perspective, it shows why score interpretation depends on understanding whether students meaningfully engaged with the tasks they received. From a transparency perspective, it demonstrates how process data can make hidden aspects of test-taking behaviour more visible to assessment developers and users. From a technology perspective, it highlights the promise of digital adaptive systems not only to measure achievement more precisely, but also to support more equitable and valid assessment experiences. The presentation will be relevant to researchers, practitioners and assessment vendors interested in designing digital assessments that are both technically strong and educationally responsible.

14:10:00 - 14:55:00

Indigenous Futures? Indigenous Continuance in an Era of Artificial Intelligence and Digital Assessment

Location: Trinity Ballroom - Salon 1

Speakers: Andrew McConnell

When educational technology and its embedded assessment systems are built without Indigenous languages,knowledge systems, or community needs, we must ask: are Indigenous futures being protected? This presentation draws on the presenter's work researching Indigenous development of AI at York University's Lassonde School of Engineering, as an Ontario educator and member of Abundant Intelligences, an international network of Indigenous community members, educators, researchers, elders and creators critically engaging with emerging technologies. The presentation examines growing Indigenous interest in AI and digital tools for language revitalization, alongside deep and warranted suspicion of corporate systems. Current large language models compress and conflate cultures, train on scraped data without community consent, and are difficult to train for accurate and useful translation. Furthermore, as Wabanakwut Kinew has observed,translation technology lacks the contextual awareness Indigenous languages require and risks interfering with the speaker relationships communities most prize. Scholars at the intersection of computational linguistics and Indigenous language rights warn further that vulnerable languages risk becoming free resources for AI development — mirroring the colonial co-option of land and resources. In response, this research explores small, open-source language models defined by and for specific community contexts, as practices of survivance, what Gerald Vizenor defines as "an active sense of presence over historical absence, the dominance of cultural simulations, and manifest manners." Crucially, where AI supports Indigenous language learning at all, the goal is ultimately to render it unnecessary, restoring the human relationships at the heart of language transmission. This has direct implications for assessment. The presentation challenges efficiency-driven standardized testing, arguing that the true measure of learning is a student's ability to recall,apply, and transform knowledge in new contexts. As a counter-model, the Uqausilirijiit Circle where Inuit elders in Ontario assess language learners through conversation and community relationship, demonstrates assessment that is rigorous, culturally grounded, and fundamentally human. The question left with the audience is not whether digital tools can assess learning, but whether any tool can know what a student needs in relation to their family, community, and future.

14:10:00 - 14:55:00

Key Themes in the 5th Edition of Educational Measurement

Location: Trinity Ballroom - Salon 4

Speakers: Dr. Mary Pitoniak

The fifth version of the book Educational Measurement was published in December, 2025. The volume has often been referred to as the “bible” of the field, making it an essential resource for researchers, practitioners,and anyone shaping the future of assessment in education. In this session, one of the co-editors of the fifth edition will describe the contents of the book, which has 21 chapters written by 52 authors. She will outline the cutting-edge research and practical guidance on educational measurement contained in the volume—from validity, reliability, and fairness to assessment design, technology, and international testing. She will also describe how the book addresses the ways in which measurement can inform teaching, learning, and accountability, as well as covering the latest technical advances and theoretical perspectives. She will emphasize key overarching themes that occur in the book, including validity, fairness, and social justice. She will also provide information on where to obtain a free electronic copy of the book. This presentation will reflect all three themes of the conference—Trust, Transparency, and Technology—and how leaders in the field view the past, present, and future of assessment in terms of these important themes.

14:10:00 - 14:55:00

When Take-Home Assessment Loses Trust: Australian Evidence on AI and Supervised Digital Exams

Location: King

Speakers: Alex Dore, Monica O'Brien

As generative AI becomes widely available to students, assessment systems face a renewed challenge: how can schools preserve fairness, authenticity and trust when substantial work can be completed outside supervised conditions? This presentation examines how generative AI is altering the trust relationship around homework and take-home assessment, and what the shift to supervised digital exams may mean for system-level assessment design. Drawing on teacher and student data from Gradeo’s 2025 report, Taking the Leap, and a forthcoming 2026 evaluation report, the paper combines student, teacher and school leader survey data with qualitative responses from schools implementing supervised digital examinations. The 2026 work explores how AI has affected trust in take-home assessment, whether students perceive peers as using AI to complete work, how well assessment after the emergence of generative AI enables students to demonstrate knowledge and skills, and which responses schools should prioritise: clearer AI policy, redesigned tasks,supervised assessment, checkpoints, viva-style questioning or other models. The presentation will discuss implications for validity, fairness, student voice, assessment integrity and public confidence. It does not argue for a narrow exam-only future. Instead, it asks which forms of assessment need supervision, which need redesign, and how technology can help build assessment systems that learners, educators and societies can trust.

14:10:00 - 14:55:00

Beyond the Marketing: Real-World Vulnerabilities in Online Proctoring

Location: Bloor

Speakers: Robert Day

Proctoring platforms promise a wide range of controls designed to protect online assessment integrity, from identity verification and audio monitoring to video surveillance and remote-control prevention. But having a feature is not the same as having an effective control. In this session, Cognisense goes beyond the marketing to examine what happens when commonly advertised proctoring capabilities are tested against real-world malpractice tactics. Drawing on assessments of the remote proctoring landscape , the session examines a number of core controls. Participants will see examples of how apparently compliant sessions can proceed despite false or implausible identification, direct verbal and AI assistance, manipulated video environments, and third-party remote control. These tactics are not theoretical or technically sophisticated. They reflect methods identified through ongoing monitoring of the assessment-integrity threat landscape and validated under controlled conditions. The findings demonstrate why organizations need to look beyond whether a platform has a security feature and ask whether that feature can demonstrably mitigate the threat it is intended to address. The session concludes with a practical integrity-assurance methodology built around assessing risk, implementing appropriate controls, testing those controls regularly, and updating the approach as threats evolve.

14:55:00 - 15:20:00
Coffee Break with Exhibitors

15:20:00 - 15:50:00
Assessment as Relationship: Community Knowledge and Culturally Responsive Practices
Location: Trinity Ballroom - Salon 4
Speakers: Aakriti Kapoor
This paper explores what it means to conceptualize educational assessment through the lens of student and community identity within the transnational schooling context of the Greater Toronto Area (GTA). As one of the most culturally diverse urban regions globally, the GTA brings together students and families who se understandings of learning, success, and accountability are shaped by varied cultural, historical, and migration experiences. Yet assessment practices in schools continue to be largely rooted in dominant Western epistemologies. The purpose of this study is to examine how culturally responsive assessment practices can enhance trust and transparency by aligning assessment with the values, identities, and priorities of students and their communities. The study draws on a qualitative literature review of case studies and research situated primarily within the Toronto District School Board (TDSB), along with other policy documents. These sources are analyzed alongside community-informed theoretical perspectives (e.g., Afrocentric frameworks, Indigenous ways of knowing). Together, these perspectives provide a critical lens to examine how assessment practices affirm or marginalize student identity, and how community knowledge systems articulate alternative understandings of educational success beyond standardized achievement measures. The review reveals competing perspectives between dominant school-based assessment practices and community-defined understandings of what matters in education. Where racialized communities in particular may prioritize holistic development, cultural continuity, relationality, and collective responsibility, institutional assessment frame works may emphasize individual performance, quantification, and comparability. When these differing value systems remain unexamined, students and families may disengage from schooling processes they perceive as culturally misaligned, while educators can experience frustration and limited capacity to meaningfully engage learners. Implications for Educational Assessment: In relation to the conference theme of Trust, Transparency,and Technology, this paper argues that culturally responsive assessment is foundational to rebuilding trust between schools and communities. Transparent assessment practices that are co-developed with communities and attentive to diverse cultural contexts can transform assessment into a tool for inclusion rather than exclusion. Ultimately, reimagining assessment requires not only new tools or technologies, but a systemic shift toward honoring community knowledge, student identity, and culturally grounded definitions of success.

15:20:00 - 15:50:00
Fostering Learning Through Assessment—A Case Study of Aga Khan University Examination Board
Location: Trinity Ballroom - Salon 5
Speakers: Dr. Naveed Yousuf, Munira Muhammad Rangwala
Improving student learning outcomes requires assessment systems to move beyond measuring achievement and focus on using assessment data to inform teaching and learning practices. This study presents a case from the Aga Khan University Examination Board (AKU-EB) in Pakistan that demonstrates how high-stakes examination data can be transformed into meaningful and actionable insights for schools. A key component of this initiative is the development and dissemination of a customized School Performance Reports (SPRs), that carries examination performance data. School Performance Report is a comprehensive document that compares each of the school's students data against the national performance. The analysis is shown at three levels: school level, subject level, and paper level. Each item is tagged with the student learning outcome and cognitive distribution. The customization enables teachers to understand their students' performance in relation to curriculum expectations, assessment objectives, and relevant benchmarks across national level. In addition,for the descriptive questions qualitative feedback is provided, allowing schools to identify strengths, learning gaps, and patterns in student achievement across subjects and competencies. By translating complex examination data into accessible insights, the reports aim to support evidence-based instructional planning and

targeted interventions. The study employed a qualitative approach, using semi-structured interviews with teachers to explore how School Performance Reports are interpreted and used within classroom practices. Teachers were asked to reflect on how SPR insights inform instructional decisions, feedback practices, and targeted support for students. Drawing on insights from the implementation of these reports, the study highlights the use of information by the teacher to reflect on teaching practices, engage in dialogue with the students, and design strategies to address targeted student learning needs. The findings demonstrate that contextualised assessment data can strengthen educators' capacity to make informed and problem-aligned pedagogical decisions and foster continuous improvement in teaching and learning. This initiative demonstrates the potential of transforming examination data into actionable insights for improving teaching and learning at the school level. It offers a promising model that can inspire other examination bodies to adopt similar practices, promoting greater use of assessment data to support targeted interventions and continuous improvement in education systems.

15:20:00 - 15:50:00

Strengthening Reading Literacy Through PIRLS - Aligned Assessment Practices in Kazakhstan

Location: King

Speakers: Diana Sartauova, Fariza Khamza

Reading literacy is a core skill that underpins students' academic success and supports lifelong learning. In the context of global educational benchmarking, international large-scale assessments such as the Progress in International Reading Literacy Study (PIRLS) play an important role in evaluating the effectiveness, transparency, and comparability of national education systems. PIRLS assesses how well fourth-grade students can retrieve information, interpret meaning, and think critically about texts, providing reliable data that can be compared across countries. Recent PIRLS results point to emerging challenges for Kazakhstan. While the country achieved a score of 536 in 2016 (27th place), this declined to 504 in 2021 (around 38th place).Although the result remains close to the international average, the downward trend raises concerns about the sustainability of reading outcomes and highlights the need for more transparent, consistent, and system-level approaches to assessment in early literacy. In response, this study presents the development and pilot implementation of PIRLS - aligned diagnostic reading assessments adapted to the national context. These tools are designed to strengthen both transparency and trust in educational assessment by providing standardized,data-informed measures of reading comprehension. The pilot was conducted in 20 schools within the Nazarbayev Intellectual Schools network, where baseline (pre-test) data were collected to identify learning gaps and enable meaningful comparisons across schools. Preliminary findings suggest that the use of standardized assessments, supported by systematic data analysis, helps build trust in assessment results and improves transparency in tracking student progress. In addition, comparative data served as a catalyst for school-level improvement, encouraging the introduction of targeted literacy practices such as structure dreading routines, guided discussions, and increased parental involvement. Overall, the study shows that technology-supported and internationally aligned assessment systems can be effective tools for improving educational quality. By enabling continuous monitoring, supporting evidence-based decision-making, and strengthening accountability, such approaches contribute to more transparent and trustworthy education systems. Improving reading literacy through reliable assessment practices not only enhances student learning outcomes but also supports the development of critical thinking, resilience, and equal opportunities for future education and careers.

15:20:00 - 15:50:00

Beyond the Rubber Stamp: Reclaiming Authentic Assessment in the Age of AI

Location: Trinity Ballroom - Salon 2

Speakers: Mark Zietara, Sandra Zietara

Is AI ruining education, or is it exposing a broken system? In this session, we move beyond "detecting cheating" to explore a radical new framework for assessment. Drawing on accelerated mastery-based learning and decentralized contribution ledgers, we will demonstrate how to decouple education from profit-centric motives and reconnect it with local community aid. Attendees will explore a "Human-in-the-Loop" architecture where AI serves as a Social Architect to manage group dynamics and trace mathematical logic. This framework compresses core academic delivery to unlock time for real-world application, while human

educators verify understanding through oral defense and peer-verified impact. Learn how to transform assessment from a private grade into a public record of contribution, making AI reliance irrelevant while restoring the "soul" of education for communal benefit.

15:20:00 - 15:50:00

The Italian Approach to Accessible and Equitable Communication of Assessment Data

Location: Bloor

Speakers: Dr. Patrizia Falzetti, Michele Cardone

Transparent and responsible communication of assessment data is essential for education systems that are committed to equity, accountability, and evidence-based decision-making. This paper presents the model developed by INVALSI to report national assessment results in a way that enhances their accessibility, clarity and interpretability for policymakers, school leaders and teachers across Italy. INVALSI administers standardised assessments in reading comprehension (Italian), mathematics, and English (reading and listening) at five key stages of the education system (grades 2, 5, 8, 10, and 13), as well as digital competencies at grades 10 and 13. The resulting large-scale datasets integrate student performance with contextual information and are transformed into structured tables, indicators, and visualisations tailored to different audiences. Schools can access interactive dashboards offering longitudinal views of results and contextualised comparisons at class, school and territorial levels. Key indicators include the distribution of proficiency levels and measures of equity such as between-class and between-school variability, as well as value-added estimates that control for prior achievement and socio-economic background. Particular attention is given to identifying critical student groups, such as implicit leavers (ILETs), who obtain an upper secondary diploma without reaching minimum competence levels. The findings demonstrate that rigorous assessment data can be communicated in a way that is both technically robust and user-friendly. Contextualised indicators and benchmarks promote fairer interpretations of results and reduce the risk of misreading raw scores. Furthermore, identifying hidden forms of disadvantage, such as that experienced by ILET students, provides a more comprehensive understanding of educational inequalities. The availability of dashboards and intuitive visual tools strengthens schools' capacity to monitor trends, identify equity gaps and respond promptly. Overall,the INVALSI approach shows how transparent, contextualised reporting can create a culture of data-informed practice. This supports targeted policy actions, equitable school comparisons and the professional development of teachers. Ultimately, it contributes to reducing disparities within the education system.

15:20:00 - 15:50:00

Designing a New Assessment Process for Transitioning Students from Primary to Secondary School in Barbados

Location: Carlton

Speakers: Stephen Scott

The transitional model used for over 40 years was a one-off test in Mathematics and English to transition from Primary to Secondary school in Barbados. However the new approach to the assessment is to have Continuous Assessment in tasks given to schools using assessments such as project-based learning, and portfolio assessments, rubrics , checklists among others. There will be four standardized tests: one each in English , Mathematics, Science and Social Studies/Civics. The tests will be more focused on application questions. A Task Force was brought together to design a different approach to transitioning from Primary School to Secondary School. After proposals were created by the Task Force the various stakeholders we reengaged to have input and buy into the various proposals. The stakeholders included focus groups of students, parents, Ministry of Education officials, teachers, Teachers Unions, principals, and experienced retired personnel. The Task Force then used the feedback to determine the preferred created outline for the new transition process and assessments. The various assessment approaches were then fully designed. Teachers and other stakeholders were trained in the new assessment methodologies and they will be coached and trained by Ministry of Education officers and personnel going forward. Town hall meetings will be held with the public and other key stakeholders inclusive of unions, Principals and teachers, and the general public to share the information. The new approach is to assist children engaging in assessment which is developmental and gives them more than one approach to be assessed in for this new approach to transition.

15:20:00 - 15:50:00

Beyond Translation Quality: Building Trustworthy, Transparent AI Support for Multilingual Assessment

Location: Bay

Speakers: César Soares

Machine translation is increasingly used in educational assessment because it promises speed, scalability, and lower costs. Yet in assessment contexts, fluent translation is not enough. A text may appear linguistically correct while still altering construct, difficulty, cultural fit, or fairness. This context creates a central challenge for assessment organizations: how to use AI to support multilingual workflows without weakening validity, reducing transparency, or losing control over sensitive materials. This presentation reports on work by the FLIP +Working Groups, an international consortium of assessment institutions building a shared multilingual item library, to develop and validate a locally hosted, auditable AI pipeline for item translation and content operations. Its purpose is to present the quality framework developed to evaluate how open-source machine translation models and Large Language Models (LLMs) can jointly support multilingual assessment work flows while strengthening trust, transparency, and data governance. The study tested a pipeline combining three open translation models (Marian MT, NLLB-200, and Translate Gemma) with automated quality checks and two local review LLM agents. The framework evaluated assessment-style items across multiple language directions, together with checks for placeholder preservation, number preservation, and output integrity. A Judge agent reviewed adequacy, fluency, and faithfulness, while a Suitability agent flagged assessment-context risks such as fairness concerns, cultural mismatch, and terminology issues. The findings show that translation quality alone is insufficient for assessment use. Linguistically acceptable outputs may still be unsuitable because of idiom failures, register shifts, terminology drift, or construct-relevant contextual changes. Translate Gemma produced the strongest average BLEU/chrF performance, but at a higher computational cost than lighter models. Placeholder and number preservation reached 100% with shielding, yet token preservation alone did not guarantee grammatically or functionally valid output. Structured review agents added value by surfacing risks that automated metrics alone could not capture. The presentation argues that responsible AI adoption in educational assessment should not rely solely on fluency metrics. Trustworthy use requires a broader quality infrastructure: transparent audit trails, automated safeguards, structured human review, and locally hosted or well-governed deployment options. This presentation offers a practical frame work for institutions seeking to apply AI in multilingual assessment while maintaining validity, security, and transparency.

16:00:00 - 16:30:00

Does the Long-Term Achievement Gap Shrink in Reading Literacy from PIRLS2001–2021?

Location: Trinity Ballroom - Salon 2

Speakers: Prof. Audrone Jakaitiene

Educational inequality, particularly the achievement gap between high- and low-performing students, is a significant concern for policymakers. This study analyses long-term trends in the reading achievement gap among 4th-grade students in 12 EU and EEA countries from 2001 to 2021, using data from the Progress in International Reading Literacy Study (PIRLS). Methods. The achievement gap is defined as the difference between the 90th and 10th percentiles of student achievement. We used multilevel growth models, the Sen’sslope, and the Mann-Kendall test to analyse trends and categorise changes in the achievement gap as favourable or unfavourable. Results. The analysis reveals that most countries are experiencing widening achievement gaps. Progress among the lowest-achieving students contributed to a reduction in gaps in 25% to75% of countries. However, declines at the 10th percentile were observed in 63% to 71% of nations, while improvements at the 90th percentile were limited. We found no significant correlation between the achievement gap and a country's average performance. Conclusion. Reading achievement inequality is increasing across these countries, indicating that policies have not effectively addressed educational disparities over the past two decades. It is crucial to explore the underlying factors driving the achievement gap, as higher-performing countries do not necessarily experience a reduction in it.

16:00:00 - 16:30:00

Ensuring Equitable and Fair Assessment: Teacher Understanding, Practices, and Tensions

Location: Trinity Ballroom - Salon 1

Speakers: Corrine Robinson, Dennis Alonzo, Dr. Karen Maras, Jonathan Ang

Equity and fairness are central yet contested principles in contemporary assessment discourse, particularly as education systems face increasing pressures for accountability, standardisation, and comparability of results. While policy frameworks frequently emphasise equitable assessment practices, there remains a limited empirical understanding of how teachers conceptualise and enact equity and fairness in everyday assessment decision-making (Rasooli et al., 2018). Existing research tends to privilege technical aspects of assessment quality, with less attention to teachers' professional judgement, interpretive work, and the contextual tensions they navigate (Phung & Michell, 2022). Moreover, equity and fairness are often treated as interchangeable constructs in policy and research, obscuring potential differences in how teachers understand and prioritise them in practice. This study addresses these gaps by examining how teachers make sense of and operationalise equity and fairness in classroom assessment. We draw on semi-structured interviews with 14teachers (with more than five years teaching experience) across diverse school contexts. The study explores four interrelated areas: teachers' conceptual understandings of equity and fairness in assessment; strategies used to ensure equitable and fair assessment design, implementation, and decision-making; tensions and conflicts encountered; and contextual and systemic factors influencing assessment practices and judgements. Data were analysed using reflexive thematic analysis to identify patterns. Findings indicate that teachers consistently distinguished between fairness, understood as the consistency and transparency of processes,and equity, understood as responsiveness to individual students' needs. Strategies used include flexible task design, multiple evidence sources, moderation, and contextualised judgement to support equitable outcomes. However, significant tensions emerged between system-level accountability expectations and teachers'commitments to responsive and inclusive assessment practices. They described navigating conflicts between standardisation and professional judgement, particularly when external reporting requirements limit opportunities to adapt assessment approaches. Influencing factors include school assessment culture,leadership expectations, access to professional learning, time constraints, and perceived trust in teacher judgement. Our findings contribute to a more nuanced understanding of equity and fairness as situated,relational, and context-dependent constructs rather than fixed technical properties of assessment. Implications include the need to reconceptualise assessment literacy to explicitly incorporate fair and equity-oriented decision-making, strengthen professional learning focused on professional judgment and interpretive expertise,and design policies that support equitable assessment practices.

16:00:00 - 16:30:00

Understanding Disengagement in TIMSS 2023 Mathematics: Evidence from Process Data and Multilevel Logistic Regression

Location: Trinity Ballroom - Salon 4

Speakers: Doris Abroampah

Student disengagement in low-stakes digital assessment threatens the validity of score interpretation because responses may reflect low effort rather than actual knowledge and skills. This study examined disengagement in Canadian TIMSS 2023 Grade 4 mathematics using process data and multilevel logistic regression. Aprocess-data-based disengagement measure was constructed from students' screen-level response times by classifying responses as rapid, idle, or effortful, and then aggregating these classifications to the student level. Rapid responses were identified using an NT20 rule, and idle responses were identified using an upper-tail outlier rule. The final multilevel analyses were based on 9,997 students nested within 579 schools. An empty random-intercept model showed that disengagement varied meaningfully across schools. The full multilevel logistic regression model indicated that confidence in mathematics and school emphasis on academic success were significant negative predictors of disengagement, whereas liking mathematics and a sense of school belonging were not significant. Robustness analyses using an alternative disengagement cutoff and a model controlling for mathematics achievement gave the same general pattern of findings. Overall, the findings suggest that disengagement in low-stakes digital mathematics assessment reflects both student-level and school-level influences and that process data can provide useful insight into low-effort responding in large-scale educational assessment.

16:00:00 - 16:30:00

Through-Year Assessment as a Summative Tool: Proof of Concept Evidence from the Pathways for Instructionally Embedded Assessment (PIE) Project

Location: Trinity Ballroom - Salon 3

Speakers: Brooke Nash, W. Jake Thompson

To increase efficiency and utility of data, large-scale assessments are increasingly expected to serve two distinct purposes: providing teachers with timely, actionable information to support student learning, and generating summative evidence of student achievement. Designing a single assessment that genuinely serves both purposes, without compromising either, remains an open challenge. The Pathways for Instructionally Embedded Assessment (PIE) project offers a proof-of-concept response to this challenge. PIE is a prototype assessment system developed and pilot-tested in 5th-grade mathematics classrooms in the Midwest region of the United States. Rather than a single end-of-year test to evaluate student achievement of grade-level learning expectations, PIE consists of short assessments throughout the school year, timed to follow instruction on groups of conceptually related content standards. Each standard is organized into a learning pathway with three progressive levels of knowledge and skills. Diagnostic classification models (DCMs) are used to score each assessment to produce skill-level mastery results that teachers can act on immediately. Evidence fromthe 2024–2025 pilot study, involving 55 teachers and approximately 1,500 students, demonstrates meaningful instructional uptake. Teachers used mastery results to identify gaps in student understanding, form targeted small groups for instruction, and adjust instruction mid-unit. Over the course of the pilot, teachers moved from initial unfamiliarity with the system toward more integrated data use, with some reporting the PIE dashboard as a primary driver of daily lesson planning by the end of the pilot study. Building on this instructional foundation,the presentation also explores a proof-of-concept for aggregating the diagnostic information accumulated across a full year of embedded assessment into a single summative result. Two psychometric approaches a redescribed—a transition DCM and a DCM/IRT hybrid model—both of which demonstrated adequate model fit and high reliability in pilot analyses, suggesting that the rich information gathered in service of instruction need not be discarded or supplemented when a year-end summary is required. Implications for the design of assessment systems that balance support for learning with the need for summative evidence will be discussed.

16:00:00 - 16:30:00

Making Assessment Decisions Visible: Transparency in Review, Scoring, and Outcomes

Location: Carlton

Speakers: Brandon Smith, Fraser Reid

Transparency in assessment is often associated with clear design and scoring criteria. However, in practice,transparency extends beyond how assessments are structured to how decisions are made, reviewed, and communicated. This session explores the role of transparency in strengthening trust in assessment systems,particularly in digital and distributed environments. It examines how decisions, such as flags, reviews, and final outcomes, can be made more visible, understandable, and explainable to stakeholders, including learners,educators, and institutions. The session will introduce a framework for transparency across the assessment life cycle, focusing on how evidence is surfaced, how decisions are documented, and how consistency is maintained. Particular attention will be given to the challenges of balancing transparency with complexity,especially when automated systems are involved. Participants will explore practical approaches to improving transparency, including clearer documentation practices, structured review processes, and methods for communicating outcomes in a way that supports trust and accountability. The session highlights the importance of not only making decisions, but making them understandable, ensuring that assessment systems can be trusted because they are visible, consistent, and explainable.

16:00:00 - 16:30:00

Exploring Artificial Intelligence Technologies' Integration for Educational Assessment in Higher Education Institutions in Nigeria

Location: Bay

Speakers: Prof. Mfonobong Umobong

The transformative potential of artificial intelligence (AI) in the educational sector is increasingly evident, particularly in the realm of assessment practices, personalized learning analytics and adaptive testing. Despite this, the scope of AI utilization in educational assessment within the higher education landscape of Nigeria

remains under explored. This study seeks to bridge this knowledge gap by investigating the adoption,effectiveness, and challenges of AI technologies in academic assessment within this context. Through a comprehensive mixed-methods approach involving quantitative surveys and qualitative interviews with 2,000faculty members across six universities in Nigeria's South-South geopolitical zone, the research delves into the awareness, usage patterns, perceived benefits, and obstacles of AI integration in educational assessment. Initial findings indicate a notable adoption rate of AI tools among lecturers yet reveal a diverse adoption landscape influenced by factors such as academic discipline, technological proficiency, and institutional support. The results highlight the potential benefits of AI in terms of improved efficiency and accuracy,personalized feedback, and reduced workload. The paper concludes with recommendations for policy makers and educational leaders to foster more effective and equitable integration of AI technologies in educational assessment. It also underscores the importance of government and institutional support as well as faculty development programs needed to bridge the gap between awareness and effective utilization of AI tools in assessment practices. Concerns regarding bias, ethical considerations, student privacy, over-reliance on standardized data, and the need for adequate training emerged as key challenges. Keywords: Artificial Intelligence, Educational Assessment, Higher Education, Technology Adoption, Personalized Feedback.

16:00:00 - 16:30:00

Using Writing Process Transparency to Ensure Academic Integrity in Digital Assessments

Location: Bloor

Speakers: Edlyn To

The rapid proliferation of generative AI has created a crisis of trust in educational assessment. Many teachers have reverted to traditional pen and paper in-class writing or defaulted to using AI detectors, which have been proven to carry high risks of false positives, especially for neurodivergent and non-native English students. Banning AI leaves students unprepared for a world where ethical use of AI and digital tools is expected. This presentation introduces an alternative solution for combating AI plagiarism developed by Canadian students and tested by thousands of teachers around the world. Aidify is a typing analytics technology that shifts the focus from AI detection to writing process transparency. By providing a detailed playback of the writing process and graphical reports of time management, educators can view the authentic evolution of a student’s writing journey including active typing analytics and all copy and paste events. Instead of relying on a probability score from a detector without any supporting evidence, this approach allows teachers to see verifiable proof of student effort. With insight into how students build their work over time rather than just looking at the finished paper, teachers’ can confidently integrate AI into the curriculum, ensuring students develop essential digital literacy skills while maintaining the highest standards of academic integrity.

16:00:00 - 16:30:00

Centering Student Voice: Building Trust Through Collaborative Assessment Design and Feedback Loops

Location: King

Speakers: Jilliam Joe, Nirupa Mathew

(Purpose) As educational landscapes evolve, building trust in assessment systems requires a fundamental shift from top-down development to collaborative design. The purpose of this presentation is to demonstrate how deliberately centering student voice and cultural relevance throughout the assessment life cycle increases student engagement, improves task efficacy, and fosters deep trust among stakeholders. (Methodology) We implemented a multi-tiered, systemic approach to integrate student feedback into our assessment architecture. First, we conducted consortium-wide student focus groups to catalog topics of authentic interest and cultural relevance. We then utilized these catalogs to train internal staff and external vendors, ensuring content creation aligns with actual student engagement. Second, during the development of new Data Science Performance Tasks (PTs) in Mathematics, we conducted classroom-based micro-pilots. These micro-pilots were specifically designed to solicit direct student feedback, allowing us to iterate and refine the PTs based on user experience before operationalizing them. Finally, we engaged student interest groups in the development of our organizational strategic plan, which culminated in establishing our primary strategic theme: developing a permanent system of continuous student feedback loops to inform, iterate, and operationalize assessment content. (Key Findings and Insights) Treating students as active co-designers yielded significant improvements

in content quality and resonance. Vendor and staff training rooted in direct student input led to the production of more responsive and engaging assessment items. During the Math PT micro-pilots, direct student critique exposed shortcomings in the task design, enabling targeted revisions that increased the tasks' effectiveness, clarity, and accessibility prior to high-stakes deployment. Furthermore, cementing this practice into our core strategic plan has shifted our organizational culture to prioritize continuous learner engagement. (Implications for Educational Assessment) Institutionalizing student voice through formal feedback loops is not merely an exercise in engagement; it is a vital mechanism for building systemic trust. By sharing our methodology and iterative outcomes, this session will provide actionable strategies for practitioners and policy-makers to transition toward collaborative assessment design. Ultimately, assessments that reflect the lived realities and authentic interests of learners are more equitable, meaningful, and trusted by the communities they serve.

16:00:00 - 16:30:00

Modelling and Predicting Academic Performance of Accounting Students in Malaysia Using Behavioural and Data-Driven Insights

Location: Trinity Ballroom - Salon 5

Speakers: Dr. Haniza Yon, Mun Yee Kok

This study examines behavioural traits associated with academic performance of accounting students in Malaysia using psychometric and machine learning approaches. A total of 280 students and alumni responded to psychometric assessments that assess their study skills, general skills and reliability. The students we redivided into two groups namely straight-pass and non-straight pass groups. Results indicated a statistically significant difference ($p<0.05$) in conscientiousness between students in the straight-pass group versus non-straight pass group, with higher conscientiousness observed among straight-pass students. This pattern is consistent with long-standing evidence from meta-analytic studies that conscientiousness is among the mos treliable and generalisable predictors of performance outcomes across contexts. Predictive modelling was performed using logistic regression and random forest approaches to evaluate multivariate relationships among behavioural traits and outcomes. The logistic regression model achieved 67% classification accuracy on a test set and identified four positive predictors namely conscientiousness, deliberation, execution, and trustworthiness. A random forest model yielded 84% accuracy when evaluated on the full dataset and highlighted the same set of contributing traits. However, the random forest performance estimate was not validated on an independent test set due to sample constraints and should be interpreted as exploratory. Overall, findings suggest that conscientiousness and execution-related competencies are consistently associated with better academic outcomes, supporting the potential use of psychometric evidence to inform targeted student development and early interventions, subject to further validation with larger samples and stronger out-of-sample evaluation. Keywords: Construct validity, Criterion-related validity, Behavioural competencies, Test development, Psychometric assessment, Academic performance, Professional accounting education, Conscientiousness, Logistic regression, Random Forest, Predictive modelling

18:00:00 - 21:30:00

Vretta Boat Cruise

Tuesday, September 29, 2026

07:00:00 - 16:30:00
Registration

07:00:00 - 08:30:00
Breakfast

08:30:00 - 09:15:00
Assessment: The Metrologist, the Messenger, and the Guide
Location: Grand Ballroom
Speakers: Thierry Rocher
We often imagine assessment as a technical problem: a matter of models, scales and precision. Thirty years of work across national and international assessment in education suggest otherwise. The real complexity of assessment is rarely where we expect it to be. This keynote follows three roles that anyone working in assessment must play. The metrologist must first be able to measure; that is the core craft. The messenger must deliver the results, even when they are unwelcome. And the guide must help others interpret those results and act on them, not simply read them. Each role reaches far beyond measurement: the choices we make are never purely technical, and never neutral. They sit in a permanent tension between methodological rigour and political and operational reality. Illustrated by experience from the French education system and international studies, the keynote argues that these challenges are shared by every education system, whatever its size or resources. And that meeting them well increasingly depends on how openly and concretely systems collaborate and learn from one another.

09:15:00 - 10:00:00
Foundational Numeracy Skills Developed in Early Years Predict Students' Mathematics Success in Senior School Years
Location: Trinity Ballroom - Salon 3
Speakers: David Cameron, Samuel Zheng, Stefanie De Jesus
The early years is a sensitive period where foundational skills of children's numerical capabilities are established, which scaffold progressively more complex mathematical competencies. It represents a critical period for early detection and intervention of children's foundational mathematic competencies. This study includes 1,538 Senior Kindergarten students from 36 schools in Toronto, who were assessed in Fall 2014,Spring 2015, and/or Spring 2016 with the Numeracy Screener developed by Western University's Brain and Mind Institute. Students who ranked below the 25th percentile in a particular task were flagged as demonstrating difficulties in the task relative to their peers. An overall at-risk indicator, which represents the number of tasks (0-5) where students scored low, and students' later performance on provincial Grade 3 (2017-2018) and Grade 9 (2023-2024) standardized mathematics assessments, were used in the analyses. Clear, stepwise relationships were observed between students' Numeracy Screener scores and their performance on both Grade 3 and Grade 9 assessments. Students who had not scored low in any Numeracy Screener task performed much better than those who scored low in one or more tasks. They also demonstrated higher rates of achievement at or above the provincial standard in both grades, and more positive mathematics self-concept and enjoyment when compared with all students in the District. Logistic regression analyses indicated that when introducing the overall at-risk indicator from the Numeracy Screener, demographic factors significantly associated with students' performance on the Grade 3 assessment were reduced to home language, parents' home presence and education level, and East Asian racial background. Similarly, when adding the early at-risk indicator, race was no longer a significant factor associated with students' performance on the Grade 9 assessment. These results show that the Numeracy Screener has the utility in identifying children earlier in their academic career for being potentially at risk for low mathematics achievement in later grades. As education systems face ongoing resource scarcity, the findings have important implications for improving system capacity for early, evidence-based interventions and instructional shifts that are aimed at developing the critical building blocks of students' early mathematics skills in support of improving their long-term achievement trajectories.

09:15:00 - 10:00:00

From Prompt to Publication: Designing Workflows That Make AI Item Generation Work

Location: Trinity Ballroom - Salon 5

Speakers: Tim Burnett

AI item generation fails when organisations treat it as a technology problem. It succeeds when they treat it as a workflow design problem. The technology is capable enough. The gap is in the systems, processes, and support structures that sit around it. Tim Burnett is the founder of the Test Community Network and has spent the past two years working directly with awarding organisations and assessment vendors on AI adoption in assessment. He led the research sprint on which this session is based, convening experienced professionals from across the industry and developing a prototype AI-assisted authoring tool to test what good workflow design looks like in practice. Organisations adopting AI for item authoring are encountering a consistent set of failure points that are structural, not technical. Without properly engineered prompts, AI produces items that are cognitively easier than intended, with weak distractors and answers that are disproportionately long. SMEs confronted with polished-looking output are susceptible to cognitive offloading, approving items with out adequate scrutiny. And when freelance SMEs work outside organisational IT infrastructure, shadow AI use becomes almost impossible to govern through policy alone. These are not problems that better training will fix. They are problems that better system design can solve. This session presents a framework built around four principles emerging from the research. Jurisdictional grounding must be built into the prompt architecture itself,specifying regulatory framework, terminology, and legal context at the point of generation rather than left to post-generation review. Retrieval-augmented generation offers a practical mechanism for domain accuracy, but source material must be strategically chunked so each generation call retrieves only what is relevant. Uploading entire qualification specifications produces expensive, diluted output. The freelance SME problem requires a provision-based response rather than restriction. Making governed, well-designed tools the path of least resistance is more effective than policies that cannot be monitored. Most fundamentally, SMEs should not be trained to prompt. They should be supported by systems that do it for them, capturing domain decisions and assembling prompts automatically so experts can focus on what they already know how to do. Attendees will leave with a reframed understanding of where AI item generation is currently going wrong, a practical framework for workflow design, and a clearer picture of what AI-assisted authoring looks like when it is built around the people who use it rather than the technology that powers it.

09:15:00 - 10:00:00

What Counts as “Good” Art in the Final Years of Schooling? Reliability and Transparency in High-Stakes Assessments in Visual Arts

Location: Trinity Ballroom - Salon 1

Speakers: Dr. Fiona Blaikie, Dr. Karen Maras

The paper compares conceptions of art assessment in the International Baccalaureate Diploma Programme (IB) and the New South Wales Stage 6 Visual Arts Syllabus (NSW). Situated against the background of standardised curricula in which high-stakes assessment and examination is prioritised, we aim to explore how the assessment of artworks produced by students in the final years of schooling is trustworthy when conducted under appropriate conditions. The paper commences with a brief account of historical conceptions of assessment in art and perceptions of the difficulty of assessing student achievement in art due to its ill-defined,or interpretive character as a knowledge domain. Following a brief overview of the IB and NSW art curricula,the structure of formative and summative assessment provisions in both contexts are outlined. We then compare how the process of assessing artworks is undertaken in each case and involve moderation, bench-marking standards, dialogic exchange and ostensive judgement, and the quality assurance of marker qualifications and training. Of interest here is how student achievement is measured against assessment criteria and outcomes that are underscored by the particular epistemological and ontological assumptions of each curriculum, factors that support transparent, objective and trustworthy assessment. In our discussion we address the impact of valid and reliable assessment practice in high-stakes centralized assessment in art onteachers’ individual and collective knowledge and understanding of what counts as ‘good’ art within each context and how this impacts student achievement. We acknowledge that while standardized assessment in art has some limitations, it also creates within the art education communities a strong consensus about what is

quality art teaching, assessment, learning and curricula. By illustrating two approaches to high-stakes testing in art, we show how commonsense assumptions about the subjective nature of assessment in art can be re-framed.

09:15:00 - 10:00:00

Digital Assessments and the “No Rush Zone” in the Upper Canada District School Board: Increasing Student and Educator Comfort as Key Components of Success

Location: Trinity Ballroom - Salon 2

Speakers: Amanda Nieman, Rob Scott

As large-scale assessments shift into digital formats, ensuring that students and educators are comfortable within these environments is critical. When unfamiliarity with digital tools or formats creates anxiety or confusion, the assessment itself can become a barrier to students accurately demonstrating their learning. This session highlights how the Upper Canada District School Board (UCDSB) has taken a deliberate, system-wide approach to EQAO that prioritizes comfort, familiarity, and confidence as essential conditions for success. Grounded in the importance of impact, precision, and alignment, UCDSB has redefined EQAO preparation by embedding EQAO-style digital assessment into everyday classroom practice. Rather than positioning EQAO as a high-stakes, isolated event, students engage in ongoing, low-stakes opportunities that mirror the digital assessment experience. Through consistent exposure to EQAO-style questions, formats, and digital tools,students develop the fluency needed to navigate the platform with ease, allowing them to focus on demonstrating their knowledge and skills. More specifically, through tools such as Knowledge hook Game Shows, interactive scavenger hunts, and the SkillScan9 D2L Platform, educators are building student confidence, familiarity, and fluency in digital environments. These strategies reflect a shift toward assessment for learning, where ongoing data informs instruction and supports timely, targeted intervention. Educator capacity is a key component of this work. System-aligned professional learning, shared resources, and cross-grade collaboration support teachers in confidently integrating EQAO-aligned practices into instruction. This coherence ensures that students experience consistent expectations and approaches across classrooms and schools. Accessibility and equity are central to UCDSB’s model. Structured supports, differentiated resources, and family engagement opportunities work together to demystify EQAO and build collective confidence. By intentionally reducing uncertainty and anxiety, the board creates conditions where all students are better positioned to succeed. Participants will leave with practical, scalable strategies for embedding EQAO-aligned practices into daily instruction, as well as a clear understanding of how increasing comfort with digital assessment can remove barriers and lead to more accurate demonstrations of student learning.

09:15:00 - 10:00:00

Evaluating the Comparability of Grade Distributions in High-Stakes National Assessments: A Post-Hoc Analysis of Hybrid Computer-Based Testing in Nigeria

Location: Carlton

Speakers: Dr. Innocent Uche Ezenwanne, Lukman Suleman, Prof. Dantani Ibrahim Wushishi

Abstract The transition from Pen and Paper Testing (PPT) to Computer-Based Examination (CBE) in high-stakes national examinations necessitates rigorous validation of score comparability to maintain stakeholder trust. This study evaluates the "mode effect" in the National Examinations Council (NECO) 2025Senior School Certificate Examination (SSCE) External in Nigeria, specifically by analysing a unique hybrid assessment model. In this modality, candidates responded to objective items digitally while viewing essay prompts on-screen, but responding via traditional paper scripts, a design intended to mitigate variances in keyboarding proficiency and digital divide constraints. Using a dataset of 95,160 candidates who sat for the NECO 2025 SSCE External, the research employs a matched-groups quasi-experimental design. A sample of4,000 candidates (n=2,000 per modality) was arrived at through stratified random sampling, ensuring parity across demographic variables and institutional types. The analysis focuses on three core subjects representing distinct cognitive domains: Mathematics (quantitative), English Language (linguistic), and Civic Education (social-evaluative). Statistical analyses, specifically the Mann-Whitney U test for ordinal grade data and Chi-Square tests of independence, were employed to assess measurement invariance and determine if the delivery format influenced the density of results within the A1–C6 (Quality Pass) range. Initial results demonstrate a high degree of comparability in the distribution of final grades across all three subjects,

suggesting that the digital interface for objective items does not introduce significant construct-irrelevant variance. Furthermore, the findings indicate that displaying essay prompts digitally while maintaining handwritten responses effectively "bridges" the digital divide. This study offers a scalable framework for digital transformation that prioritises transparency and validity in resource-constrained environments. By providing empirical evidence from a live, large-scale assessment, the research contributes to the global discourse on maintaining assessment integrity during technological transitions. Keywords: Hybrid Computer-Based Testing(CBE), Mode-Effect, Grade Distribution, Mann-Whitney U, Measurement Invariance, Educational Equity,Nigeria, High-Stakes Assessment.

09:15:00 - 10:00:00

Reconstructing Validity in the Age of AI: A Stage-Differentiated Framework for K–12 Assessment

Location: Bay

Speakers: Dr. Kathryn Richardson

The rapid integration of artificial intelligence (AI) into educational contexts has prompted significant debate regarding academic integrity in K–12 settings. However, framing AI primarily as a threat to honesty misidentifies the deeper problem which is that AI redistributes cognition across the assessment cycle, fundamentally challenging whether assessments continue to measure what they claim to measure and fulfil their intended purpose. The critical question is not whether students cheat, but whether assessment retains construct validity and fitness for purpose when AI is available to all actors at every stage of the assessment process. Drawing onMessick's (1989) unified construct validity framework, Kane's (2006) argument-based approach to validity, andNewton's (2007, 2014) concept of fitness for purpose, this paper argues that AI introduces differential threats to validity depending on the stage of the assessment cycle at which it is deployed and the actor deploying it. To date, existing frameworks assess AI use holistically across assessments without disaggregating by stage or actor. This paper contends that such approaches are theoretically insufficient for identifying where, how, and by whom validity is disrupted. Employing a critical literature review methodology, this paper systematically examines literature spanning validity theory, assessment design, and AI integration in K–12 education. There view identifies a critical gap: no existing framework disaggregates AI-related validity threats by assessment stage or actor. The paper maps the assessment cycle across two actor perspectives, educator and student,identifying discrete stages at which AI engagement threatens the inferential chain from observed performance to score interpretation to use. The paper proposes a stage-differentiated validity framework for AI use in K–12assessment, offering recommendations for assessment designers, practitioners, and policymakers seeking top reserve construct integrity and fitness for purpose in AI-saturated educational environments.

09:15:00 - 10:00:00

Improving Educator Understanding of Students via Collaborative Scoring ofStudent Work

Location: King

Speakers: Dr. Edward Roeber, Kenna Fries

It's ironic in this day of using artificial intelligence (AI) to evaluate large-scale assessment responses to suggest that educators hand-score student essays or performance assessment responses. However, this can be a valuable means for teachers to gain insights into students' level of achievement and suggest how to move student learning forward. It helps teachers understand what students achieved, promotes understanding of why and what to do next, and reveals students' thinking processes and misconceptions. This gives teachers a richer picture of student achievement than automated essay grading. Teachers can also see a broader range of student achievement, especially useful for teachers with classes of low-achieving students, to assure that the irinstructional and grading practices support rigorous outcomes. If such professional learning has such benefits,why isn't it used more often? Often, it's because of logistics involved. Bringing teachers together involves meeting and travel expenses, and providing tools to view and score student responses. This is especially truefor provincial or state common assessments. Pearson Assessment's Online Scoring and Reporting (OSCAR)software addresses these issues. It permits educators to individually score assessments remotely, avoiding logistics costs without sacrificing the benefits of collaborative scoring. Teacher scoring of student work provides unbeatable insights, leading to improved student learning and achievement. Training to use OSCAR provides a consistent, coherent view of how to apply rubrics so that teachers understand the range of performances across the score scale and each score point. During training, teachers also understand

where they agree and disagree to further understand how to apply the rubric. Finally, once teachers start scoring, they can see the range of writing as an excellent basis for discussion and professional development. The OSCAR software might also be used in these ways: Student Assessment Uses Permit districts to construct and use their own interim assessments and measure changes in student performance. Educator Uses Permit educators to assemble evidence of effectiveness in individual Personal Professional Practices portfolios. Enable educator review of exemplary teachers' individual Personal Professional Practices portfolios. Potential provincial/state and national uses Enable districts to compare student achievement within the same region, province/state, or nation.

09:15:00 - 10:00:00

Beyond Difficulty - Next-Generation Adaptive Assessment That Measures How Learners Think, Not Just How Well They Score

Location: Trinity Ballroom - Salon 4

Speakers: Sanketa Anand

Computerized Adaptive Testing has existed for decades — and for most of that time, it has done one thing:select the next item based on a single estimated ability score. The candidate answers; the system recalibrates difficulty; the next item appears. This is adaptive testing as a logistics optimization — more efficient than fixed-form delivery, but no more insightful about the learner sitting behind the screen. The next generation of adaptive assessment systems is doing something fundamentally different. Drawing on reinforcement learning,multidimensional item response theory, and real-time process data — response latency, revision behaviour, navigation patterns, and error trajectories — these systems are beginning to model not just how well a learner performs, but how they think. They distinguish between candidates who arrive at correct answers confidently and those who arrive there through elimination and guesswork. They detect fatigue, disengagement, and overconfidence. They adapt not only item difficulty but item format, cognitive demand, and pacing in response to inferred cognitive state. This presentation examines what this shift means for educational assessment in practice. It begins with an honest account of classical CAT — its genuine strengths in measurement efficiency and its limitations as a window into learner cognition. It then examines the evidence base for next-generation multidimensional adaptive systems: where the psychometric validation supports their use, where it does not yet, and what implementation prerequisites look like in terms of item bank depth, infrastructure, and scoring model complexity. Critically, the session addresses the trust and transparency dimensions that multidimensional adaptive assessment raises. If the system is modelling cognitive processes, not just ability,candidates and institutions deserve to understand what is being inferred, how those inferences affect scores,and what safeguards prevent adaptive algorithms from encoding or amplifying demographic bias.

10:00:00 - 10:25:00

Coffee Break with Exhibitors

10:25:00 - 11:25:00

From the Lab to the Classroom: Collaboration Among Researchers, Educators, and Industry Partners in the Development of a Digital Numeracy Screener

Location: Grand Ballroom

Speakers: Charles Anifowose, Dr. Heather Douglas, Jo-Anne LeFevre, Michael Slipenkyj, Ron Cadez

Ensuring trust and transparency when technology is used to develop assessments requires a team approach. Researchers and educators are the "subject matter experts" who provide the content and the oversight to ensure that assessment tools are evidence based, valid, and useful for the intended purpose. They must work together with technologists to ensure quality products. Researchers from Carleton University and education partners from school divisions across Canada are collaborating with technologists from Vretta. on a digital version of the Early Math Assessment at School (EMA@School). The EMA@School is a traditional paper-based screening tool. It is used to identify students needing additional support, for progress monitoring,to measure students' growth in numeracy in response to interventions, and to inform school division policies and budget decisions. In this panel, we bring together an educator (Cadez), a researcher (Douglas), and an industry partner (Anifowose), to discuss their experiences working together to translate the paper-based EMA@School to the Digital EMA@School. Panel members will each give a brief introduction, explaining

their role in the project. Next, they will respond to a series of prompts addressing the core issues of trust and transparency in the development of this digital screener. Each of the participants in this process have unique and important roles to play in ensuring that there is transparency in screener design and scoring. Collaboration is critical to ensure equity and fairness across student populations, support appropriate data sharing, and develop reporting conventions that will maximize the benefits of the digital tool. Each panel member will respond to prompts designed to probe the benefits and challenges they have experienced working on this project. The audience will be encouraged to ask questions after each prompt. An example prompt is: “What key advice would you give to someone in your role who is contemplating becoming involved in a similar project?” Implications for Educational Assessment: The collaborative process, involving contributors from the lab to the classroom, supports the goal of ensuring that the final version of the digital screener will be evidence-based, reliable, valid, and genuinely useful for educators.

11:30:00 - 12:00:00

Conceptualisations of Culturally Responsive Assessment: A PRISMA Systematic Review

Location: Trinity Ballroom - Salon 1

Speakers: Corrine Robinson, Dennis Alonzo, Dr. Karen Maras, Jonathan Ang

Culturally responsive assessment (CRA) has gained increasing prominence as education systems seek to address persistent inequities in student outcomes and to recognise the cultural, linguistic, and epistemic diversity of learners (Nortvedt et al., 2020). CRA is frequently positioned as a mechanism for strengthening both equity and validity in assessment. However, despite growing scholarly attention, the field remains conceptually fragmented, with inconsistent definitions, competing theoretical foundations, and uneven levels of empirical specificity (Herzog-Punzenberger et al., 2020). Much of the existing literature foregrounds culturally responsive pedagogy, with assessment treated as a secondary or implicit component. Furthermore, CRA scholarship often remains disconnected from mainstream assessment theory, particularly validity, reliability, and fairness frameworks, limiting its uptake in policy and practice. There is also limited synthesis of how CRA is operationalised across contexts. This study addresses these gaps by systematically examining how CRA is conceptualised, operationalised, and evidenced in education research. Guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses methodology, a search was conducted across ERIC, Scopus, and Web of Science. Peer-reviewed journal articles published between 2000 and 2025 were identified using predefined search terms. Article screening was conducted in accordance with established inclusion and exclusion criteria. A final corpus of 112 studies was included. Data were analysed using qualitative thematic synthesis to identify dominant conceptual patterns and areas of conceptual and methodological convergence and divergence. Findings reveal four dominant conceptualisations of CRA: Fairness and bias mitigation approaches focused on reducing cultural bias in assessment design and scoring.Pedagogical differentiation approaches emphasising adaptation of assessment tasks to student cultural and linguistic backgrounds. Participatory and co-constructed assessment approaches foregrounding student voice and community knowledge. Critical socio-political approaches positioning assessment as a site for challenging systemic inequities. Across these strands, CRA was frequently invoked rhetorically, with limited operational clarity, minimal empirical testing of outcomes, and weak integration with assessment theory and frameworks. Findings highlight the need for stronger conceptual coherence, explicit alignment between CRA and contemporary validity theory, and more rigorous empirical evidence. Implications include developing integrative CRA assessment frameworks, refining assessment literacy to include cultural responsiveness as a core competence, and providing policy guidance that supports culturally responsive assessment.

11:30:00 - 12:00:00

Evaluating Consistency and Fairness Between the November and June National Senior Certificate Examinations

Location: Trinity Ballroom - Salon 4

Speakers: Doctor Phokwani

The introduction of the May/June National Senior Certificate (NSC) examination in South Africa as a fully-fledged assessment opportunity necessitates rigorous evaluation of its consistency and fairness relative to the traditional October/November examination. Given the high-stakes nature of the NSC, it is essential that both examination cycles measure equivalent competencies and yield comparable outcomes to ensure reliability,

credibility, and public confidence in the assessment system. This study evaluates the consistency, fairness, and equivalence of the October/November and May/June NSC examinations across multiple subjects over a four-year period. Using a quantitative research design, the study analysed secondary performance data of candidates who wrote both examination cycles between 2021 and 2024. The analysis focused on high-enrolment subjects administered by the one of the two private assessment bodies offering the NSC. Statistical techniques included the Coefficient of Stability and Equivalence (COSE), operationalised through Pearson's correlation coefficient, Lin's Concordance Correlation Coefficient (CCC), and linear regression analysis to assess stability, agreement, and predictive relationships between examination scores. The findings indicate acceptable to excellent stability for nine of the eleven subjects analysed, with COSE values generally exceeding established reliability thresholds. Regression analyses further demonstrate strong predictive relationships, with a substantial proportion of the variance in May/June examination marks explained by October/November performance, providing empirical evidence of mark equivalence across examination cycles. While Tourism and Economics displayed limited stability in isolated years, overall trends for these subjects remained acceptable across the broader period of analysis. The study concludes that the October/November and May/June NSC examinations are largely consistent, fair, and equivalent, supporting the legitimacy of the May/June examination as an alternative assessment opportunity. These findings contribute to the discourse on assessment reliability and fairness and provide an empirical model for evaluating equivalence across examination administrations within high-stakes assessment systems. Keywords: consistency, fairness, equivalence, National Senior Certificate (NSC), May/June examination, October/November examination, assessment reliability, COSE, Pearson correlation, Lin's concordance correlation coefficient, regression analysis

11:30:00 - 12:00:00

Navigating the Landscape of Multiple Data Sources for Educational Reporting in LMICs: The Case of Pakistan

Location: Trinity Ballroom - Salon 2

Speakers: Saima Maheen

This paper offers an experience-informed examination of Pakistan's fragmented assessment ecosystem, drawing on the authors' involvement in designing and implementing provincial assessment systems and systematic document analysis. Competing agencies, divergent methodologies, and misaligned purposes impede the effective use of assessment data for reporting and improving learning outcomes. Pakistan's assessment ecosystem has experienced substantial expansion over the last two decades. Provincial agencies,[1] citizen-led bodies,[2] and specialized organizations[3] now produce learning assessment data. Differences in their procedures and purposes yield divergent outcomes within the same context, leading to contested claims about learning outcomes and eroding trust. The question of which data to trust, and for what purpose, has become structurally unresolved. A central analytic lens is the political economy of assessment data: how institutional interests, accountability, and resource dependencies shape measurement and reporting, and whose data is authoritative and correct. International reporting frameworks, such as SDG 4, embed their own standards, excluding subnational and non-state data,[4] while creating incentives to produce data for external audiences rather than system use. Concurrently, advocacy-driven citizen-led assessments are routinely deployed for purposes that exceed their methodological design, raising concerns about interpretation and the interests served. The result is a decoupling of data production from accountability and improvement. The paper identifies two structural dimensions of this challenge: 1) outward-facing: international and national reporting obligations are met, yet the data is fragmented, unevenly legitimate, and methodologically incompatible; 2) inward-facing: assessment results have limited influence on system-level decisions — curriculum reform, teacher development, financial allocation — and the sporadic data use depends on externally funded technical assistance rather than institutionalized capacity. These dimensions are interdependent; the prioritization of reporting for external audiences has displaced investment in, and the incentives for, institutional infrastructure required for evidence-informed practice. The paper proposes a three-pronged governance framework: clarifying the purpose and authority of each assessment actor; restructuring coordination mechanisms to reduce duplication and align reporting with use; and building institutionalized data literacy and utilization incentives at the system level. These insights have implications for systems seeking to move beyond data production toward assessment ecosystems that are transparent in their

reporting and consequential in their practice. References: Punjab Examination Commission, Government of Punjab ASER Pakistan. https://aserpakistan.org/report National Testing Service, Pakistan UNESCO. Institute for Statistics (2018)

11:30:00 - 12:00:00

Designing for Transparency: Making Assessment Design and Scoring Explicit in High-Stakes Examinations

Location: Trinity Ballroom - Salon 5

Speakers: Robyn Mowatt

Transparency in assessment design and scoring is critical to the credibility of high-stakes examination systems, particularly where stakeholders require clear evidence of how standards are defined and applied. Using a case study approach, this paper analyses subject assessment guidelines, moderation tools, and qualitative and quantitative reports from the Independent Examinations Board's National Senior Certificate (NSC). It argues that transparency can serve as a deliberate design strategy to support validity, reliability, and fairness while maintaining stakeholder confidence. In many contexts, credibility is supported through technical processes such as pretesting and statistical equating; however, in the NSC, these approaches are not currently used at scale or required by regulation. The paper examines three IEB interventions that make assessment processes explicit to schools: Firstly, a common grid, used as a design grid during paper setting and as an analysis grid during moderation, specifies, at question level, the cognitive taxonomy and, where applicable, the level of difficulty, together with the distribution of marks across levels and the scope of content coverage. Released to schools after each examination sitting alongside the question paper and marking guidelines, these grids make cognitive demand, difficulty level, and content coverage explicit, establishing a common professional understanding of standards across the system. Secondly, Itemised Data Capture (IDC) reports provide schools with raw cohort-level performance data alongside school-level results per question, as well as mean, median, and standard deviation, enabling analysis of learner performance and providing visibility into scoring patterns and any subject-level adjustments applied during standardisation. Thirdly, statistical moderation of school-based assessment (SBA), which constitutes 25% of the NSC and is conducted internally by schools, is made explicit through reporting. The rules governing the statistical transformation of SBA marks to align with examination performance are transparent, enabling schools to see how internal assessment outcomes are adjusted relative to external standards. Combined with qualitative feedback from trained moderators, these reports make changes to SBA marks visible and support schools in monitoring and strengthening their internal assessment standards, thereby enhancing accountability at school level. Taken together, these design features demonstrate how transparency in both assessment design and scoring strengthens the defensibility of examination outcomes while building assessment literacy among teachers and schools. By making cognitive demand, assessment criteria, and scoring decisions visible, the system supports a shared understanding of standards, with implications for high-stakes contexts seeking to enhance credibility through design. Keywords: assessment design; data-informed practice; validity; reliability; capacity building

11:30:00 - 12:00:00

Equity in Assessments Across Different Populations and Contexts: The INVALSI Experience in Italy

Location: Bay

Speakers: Dr. Patrizia Falzetti, Michele Cardone

It is essential to ensure fairness and impartiality in large-scale assessments, allowing all students, regardless of their socio-economic status, cultural or linguistic background, gender or individual characteristics, to demonstrate their competencies in an authentic and comparable way. This paper presents INVALSI's experience of addressing equity challenges within the Italian national assessment system, which involves almost 2.5 million students annually from diverse social, geographical and linguistic contexts. INVALSI has developed a technical framework designed to promote equity throughout the entire assessment process, from item development to test administration. Core procedures include systematic pre-testing, differential item functioning (DIF) analysis, scaling and iterative item review. These procedures ensure that test items function consistently across population subgroups, such as males and females, native and non-native Italian speakers, and students from different socio-economic backgrounds. The transition to computer-based testing has further strengthened the standardisation of administration conditions while enabling adjustments for students with

special educational needs. Longitudinal analyses demonstrate that this multi-layered approach significantly reduces sources of measurement bias and enhances the comparability of assessment results across heterogeneous student populations. Additionally, integrating contextualised indicators — such as the ESCS index, value-added measures, and indicators of variability between classes and schools — allows for detailed monitoring of inequalities across regions, schools, and educational settings. These indicators are made available to schools through dedicated reporting platforms and to the education system as a whole at the macro-policy level. Overall, the INVALSI experience demonstrates how a national assessment system can successfully combine methodological rigour with a strong commitment to equity. As well as improving the validity and fairness of assessments, this approach provides policymakers, schools and researchers with robust tools to identify disparities in educational opportunities and design targeted interventions to promote equity throughout the education system.

11:30:00 - 12:00:00

Exploring Technology Use for Greater Inclusion in the Caribbean Examinations Council's Assessments

Location: Trinity Ballroom - Salon 3

Speakers: James Young

The Caribbean Examinations Council (CXC®) remains committed to ensuring that its assessments are equitable, accessible, and responsive to the diverse needs of learners across the region, while upholding measurement principles such as fairness and validity. As global digital transformation accelerates, technology presents important opportunities to enhance inclusion in educational assessment. CXC® has long offered Special Assessment Arrangements (SAA) for candidates with varied challenges, including screen reader-compatible papers, enlarged print formats, and provision of readers and writers. While effective, these arrangements are limited within paper-based systems. Emerging from the 2026 Inaugural CXC® Conference was the need to leverage available technology to broaden accessibility, streamline administration, and ensure more consistent support at scale. This paper examines how technology-enabled approaches—from digital test delivery to assistive tools—can reduce barriers for visually-impaired candidates, candidates with dyslexia, and candidates who have physical ailments that inhibit their writing ability. It specifically explores CXC's efforts to enhance SAA through technology, with particular focus on assistive tools and speech capture methods for recording candidate responses. The research is conducted in two phases: Phase One comprises a desk review of accommodation requirements and expert evaluation of the testing platform for fitness for purpose. Findings highlight areas for improved accessibility, including font options, background colour adjustments, and enhanced compatibility with screen reader protocols. Phase Two is grounded in a qualitative approach, utilizing purposive sampling to examine candidate interactions across selected territories with technology-enabled assessments. It focuses on (i) platform accessibility experiences and (ii) the utility of capturing oral responses. The study aims to identify patterns of disadvantage as well as inform refinements to item design and assessment development and administration. Additionally, the results can guide professional development for educators in accommodating special assessments. When thoughtfully deployed and supported by inclusive policies, technology can serve as a powerful catalyst for equity within CXC® assessments. By prioritizing accessibility and adopting a regionally responsive approach, CXC® can build an assessment ecosystem that reflects the diversity of Caribbean learners and supports continued educational advancement. The findings will contribute to ongoing discussions on the future of assessment and will help inform CXC's strategic direction as it deepens its digital transformation.

11:30:00 - 12:00:00

Curriculum Expectation-Focused Assessment and Evaluation in the LDCSB

Location: Bloor

Speakers: Steven Floyd

In September 2024, the London District Catholic School Board (LCDSB) implemented changes to their Secondary (grade 9-12) Assessment and Evaluation procedures across all subject areas in their school board's nine high schools. The new procedures focused on transparency and clarity while maintaining alignment with provincial policies and standards. A careful and intentional implementation plan was followed that revolved around the three main pillars of what evaluations look like to students, how student performance can be tracked, and how final reports can be developed. Two years into these changes, the school board is

recognizing newfound opportunities resulting from the change, including facilitated fidelity to course expectations as well as improved alignment in data collection. In this session presentation, the LDCSB's journey through these changes will be presented, from initial information gathering and stakeholder input, to communication and implementation strategies and challenges, and to the new opportunities now afforded to students, teachers, and the school board as a result of the change.

11:30:00 - 12:00:00

Uneven Access, Emerging Practices: Technology-Enhanced English Assessment in Asian Higher Education

Location: King

Speakers: Dr. Saraswati Dawadi

The primary objective of this presentation is to report findings from a research project examining the use of digital and mobile technology (DMT) and generative artificial intelligence (GenAI) in the teaching, learning, and assessment of English (TALE) in higher education across Bangladesh, China, India, and Indonesia, the four most populous countries in East and South Asia. Drawing on data collected from nearly 6,000 students and more than 300 teachers through a mixed methods design incorporating online surveys, focus group discussions, Padlet activities, and semi structured interviews, the study identifies significant disparities in students' access to DMT, prevailing practices in the use of DMT and GenAI for English assessment, and the role of technology in promoting or hindering inclusion in higher education. The findings further indicate that, compared with their use in English teaching and learning, the application of DMT and GenAI in assessment is considerably less common, largely due to limited teacher expertise in technology enhanced assessment. These insights carry important implications for both policy and practice.

12:00:00 - 13:00:00

Lunch

13:00:00 - 13:30:00

Reducing Teacher Workload with Evidence?Led AI

Location: Grand Ballroom

Speakers: Jim Crawford

This session demonstrates how a number of schools in the UK worked collaboratively with assessment technology specialists, Surpass, to introduce an AI-supported marking platform that accurately transcribes lengthy handwritten or typed practice exam scripts, assigns marks against a calibrated marking scheme/scoring rubric, and generates feedback for individual learners and teachers - with teacher in the loop controls at every stage. The audience will learn about: The specific challenges faced by teachers and school leaders in the UK, who are required to mark and provide feedback on essay-heavy subjects such as English Literature, English Language and History across the 14-18 age groups, as part of their school leaving exam preparation. The calibration process involved line-by-line comparison between AI outputs and expert judgements, with teachers double-marking all scripts to verify accuracy. The results of the study, which showed:Traditional marking that consumes approximately 50 hours per teacher was reduced to under an hour. Feedback was significantly improved for individual learners, teachers and school leaders. AI marking was consistently within 2 marks of a Senior Examiner on 80-mark papers (compared to ±8 mark tolerance between human examiners). The session will also demonstrate the final solution, which provides teachers with a simple interface to digitise handwritten scripts using OCR, before marking each student response against expert-calibrated criteria and automatically producing personalised feedback. The study is based on over 2,000 student scripts across English and History subjects, and our evidence comes from real qualitative and quantitative data: "If we had only used this solution earlier, my teachers would have got three weekends back." (Head Teacher) Serious AI marking solutions require expert calibration and collaboration; industry experts acknowledge this. This study reflects how responsible AI should develop: carefully, with evidence, and in partnership with schools.

13:30:00 - 14:00:00

Responsible Use of Artificial Intelligence Across the Assessment Lifecycle: Governance Led Practice from Large Scale Assessment Delivery

Location: Grand Ballroom

Speakers: Dr. Isabelle Gonthier

As Artificial Intelligence (AI) becomes increasingly embedded in educational assessment, assessment organizations face growing expectations to demonstrate that its use is transparent, well governed, and aligned with the evidentiary standards of high stakes testing. This paper contributes to the assessment community by sharing governance led practice from a global assessment organization, using a representative large scale assessment program as an illustrative example of how AI can be applied responsibly across assessment development, scoring quality assurance, and test security. Within assessment development, AI is used in narrowly defined, supporting roles to assist controlled item generation and content analysis. These activities are embedded within an established test development lifecycle that includes construct definition, expert authoring, structured review, pretesting, and formal approval prior to operational use. AI supported content is reviewed, edited, and authorized by qualified subject matter experts, and no unreviewed or solely AI generated content is introduced directly into live testing. In scoring, AI based rating technologies are applied to support human rating through structured quality assurance processes for constructed response (CR) items. AI outputs are used to evaluate consistency, detect anomalous patterns, and support targeted escalation. Human raters retain responsibility for assigning final scores, with defined procedures for second rating and adjudication when discrepancies arise between human and AI ratings or among human raters. This hybrid approach supports reliability and scalability while preserving professional judgement. In test security, AI enabled biometric matching, plagiarism detection, and integrity analytics are used to support identity assurance and to identify patterns requiring investigation before, during, and after test administration. All AI generated alerts are reviewed through documented human adjudication processes, and no outcome action is taken without human validation. Across all use cases, AI deployment is governed by formal approval processes, risk-based review, and ongoing oversight. Model evaluation, bias assessment, drift monitoring, incident management, data protection, and accessibility reviews are integrated into governance and operational workflows. AI systems do not autonomously determine final scores, cancel results, or make binding eligibility related decisions. By articulating this end-to-end, governance first approach, the paper aims to support collective learning and contribute practical, transferable guidance on how AI can be leveraged to enhance efficiency and analytical insight while maintaining validity, fairness, accountability, and public trust in assessment outcomes. The framework demonstrates how provisional parameter predictions can support earlier form assembly, reduce reliance on large-scale item tryouts, and improve efficiency while preserving transparency.

14:10:00 - 14:55:00

Evaluating the Transformative Potential of AI-Driven Item Generation for Chartered Insurance Professional Examination

Location: Trinity Ballroom - Salon 5

Speakers: Aida Hadziomerovic, Gelareh Jowkar, Sanketa Anand

In this joint session, Excelsoft Technologies and the Insurance Institute of Canada (IIC) share insights from a pilot that reimagines assessment development through AI-powered item generation for the Chartered Insurance Professional (CIP) program. Results of this pilot project, key performance indicators evaluated, and areas of strength and limitations will be discussed. During this pilot, 537 multiple-choice and constructed-response items were generated using AI models trained on course-specific content aligned to defined learning objectives, for use on Chartered Insurance Professional (CIP) examinations. By combining Bloom's taxonomy levels (Remember, Understand, Apply) with controlled creativity parameters (Moderate, High, Very High), the approach enabled systematic generation of items across cognitive levels. The combination of taxonomy and creativity metrics allowed for creation of questions with varying degrees of adherence to course content. Notably, higher creativity settings proved effective in producing questions that assess deeper understanding and application; highlighting AI's potential to move beyond basic, knowledge-based item generation, toward higher-order assessment design. The questions were evaluated by subject matter experts on format, accuracy and content validity. Question quality was rated by the extent of revision required, with 70% of questions rated as requiring no or minimal revision. Variations observed on this

rating based on factors such as question type, taxonomy classification and creativity, provide further insight into how AI can be optimally configured for different assessment needs. Beyond efficiency gains, this pilot positions AI as a catalyst for transforming item development; enabling scalable, consistent, and pedagogically aligned assessment design. The session will explore practical implementation considerations, and the evolving role of AI in shaping the future of high-stakes assessments.

14:10:00 - 14:55:00
Math Proficiency Test Preparation: Digital, Accessible, Adaptive Assessments of Knowledge
Location: Bay
Speakers: Dr. Lisa Briona, Isha DeCoito
Traditionally, assessment practices relied heavily upon students' rote memorization of facts and figures and failed to ascertain students' conceptual understanding. Assessment is witnessing a leap in terms of innovative technology, such as the use of artificial intelligence in grading, as well as the use of digital assessments that capture students' skills and/or conceptual understanding. Dedicated digital assessment can improve fairness and reduce instructors' unconscious biases that can negatively affect learning outcomes for certain groups of students based on gender, race, socioeconomic status, ability, and various other identifying characteristics. An example of such technology was demonstrated in a prototype of a digital assessment repository and successfully tested in an undergraduate science class. Building upon lessons learned, the revised Digital, Accessible, Adaptive Assessments of Knowledge (DA3K) platform was refined and expanded to support real-time differentiated assessment to better ascertain student understanding. The principles of EDID are embedded in DA3K – to provide students with an accessible digital assessment, the design of DA3K follows accessibility regulations and support the use of assistive technologies such as screen readers, speech input software, and others. Through DA3K, we aim to Identify and potentially eliminate bias in assessment strategies for equity deserving groups (including women, minoritized and racialized groups, 2SLGBTQIA+, persons with disabilities, and Indigenous peoples) Support differentiated modes of learning. Our presentation focuses on teacher candidates' (TCs') engagement with DA3K as they prepare for Ontario's Math Proficiency Test (MPT), a requirement for certification with the Ontario College of Teachers. Highlights will include DA3K's ongoing support of TCs in the form of diagnostic, formative and summative assessments as TCs prepare for the MPT, including challenges and identified areas for improvement in teacher education programs.

14:10:00 - 14:55:00
Call and Response: How a Large-Scale Assessment Organization Prioritized Indigenous Reconciliation
Location: Trinity Ballroom - Salon 1
Speakers: Derek Chen, Katie Wilhelm, V West
As large-scale assessment programs grapple with the legacy of dominant-culture bias, strategies to increase equity and fairness in testing are paramount. In the 2021–2022 Annual Report, EQAO outlined three strategic priorities. Strategic Priority 2 states, "Commit to action on all applicable Truth and Reconciliation Commission [of Canada (TRC)] recommendations and the development of an agency-wide equity, diversity and inclusion plan." This presentation offers assessment professionals a practical case study in enhancing fairness and accessibility for Indigenous students during the development, review and implementation of large-scale assessments. The presenters will describe strategies used at EQAO including: Creating a senior-executive-level position with a focus on equity; Establishing partnerships with members of Indigenous communities; Providing professional-learning sessions and activities to staff; Crafting training materials for assessment developers and reviewers designed to raise cultural awareness and responsiveness for Indigenous test-takers. Key outcomes included: Increase in assessment items intended to create enhanced access for Indigenous students; Data showing that professional-learning sessions increased employee awareness of accessibility barriers present in the organization's work Increased environment of awareness in the workplace Increase in comments from students who observed feeling affirmed and engaged by the assessment content. The initiative demonstrated how responding to recommendations from identified communities and applying best practices from the education-measurement field can result in enhanced fairness and cultural relevance in both the assessment and the workplace. Our session includes methodologies meant to disrupt the belief that culturally "neutral" assessments are unbiased. These include strategies to: Educate content developers and reviewers about the prevalence of dominant-culture bias in assessments Provide professional-learning sessions and activities to raise staff awareness of barriers to equity and fairness faced by Indigenous students and communities Implement guideline principles for the creation of culturally relevant and responsive assessment items Employ Indigenous writers, reviewers and advisors skilled in creating items with culturally appropriate content.

14:10:00 - 14:55:00

From a Messy Testing Patchwork to Coherent Systems of High-Quality Instructional Assessments in the United States

Location: Trinity Ballroom - Salon 2

Speakers: Jennifer Vranek, Michelle Odemwingie

Despite decades of education reform in the United States, student achievement in core academic subjects is stagnant or declining nationwide. Eager to improve results, policymakers have increased assessment requirements, resulting in a messy patchwork of federal, state and local requirements. Education First's study of K–8 assessments and state policies across 42 U.S. states, Rethinking the Test Pile, is the first in-depth look in more than a decade about how much student testing is actually happening in American schools. We examined what academic assessments students take and why, how much time and money they require, how leaders decide what to administer and how state policy choices shape local systems. We catalogued assessments, interviewed district and state leaders, analyzed assessment statutes and guidance and reviewed past audit and assessment literacy initiatives. The research indicates that American students take upwards of 80 different standardized assessments in kindergarten through eighth grade--not counting curriculum-embedded assessments or formative assessments. Despite all this testing, we found no statistically significant relationship between assessment volume and proficiency on English or mathematics summative assessments, nor growth over time in English or math proficiency. At the same time, Achievement Network studied schools that are making rapid gains in improving student learning. In new research synthesized in The Missing Link, Achievement Network argues for a new conception of local assessments that have instructional utility, curricular integrity and data-driven continuity. As U.S. federal leadership for education recedes, it's up to states and school systems to streamline assessments and strengthen coherence across state standards,curricular materials and assessments. What makes an assessment truly instructionally useful? This interactive session draws on Education First's and Achievement Network's research to explore how assessment design and use can either undermine or reinforce the implementation of evidence-based practices, including high-quality instructional materials (HQIM). Participants will examine practical frameworks, real-world examples, and reflection tools to assess the instructional utility of their own systems—leaving with strategies to better align assessments, curriculum and teaching for stronger, more coherent learning experiences. This session is designed for system leaders, policy advocates and others who influence or guide assessment and curricular policy.

14:10:00 - 14:55:00

From Print to Pixels: Reading Literacy in the Digital Age

Location: Carlton

Speakers: Jeneve Swaby

The integration of computers and digital technologies into everyday life has become ubiquitous, reshaping how people learn, work, communicate, and access information. In turn, educational assessment has increasingly moved into digital environments that more closely reflect the contexts in which students now read and learn. Digital assessments offer the added advantage of capturing not only achievement outcomes but also process data that reveal how students interact with tasks. At the same time, despite benefits such as efficiency and support for innovative item types, digital assessment raises important questions about construct representation,accessibility, and score interpretation. Using data from the Progress in International Reading Literacy Study (PIRLS), an international large-scale assessment of fourth-grade reading literacy, these studies address two overarching questions: What do digital reading assessments measure?How do contextual and behavioral factors shape performance in digital reading environments? The first study used linked data from PIRLS 2016and electronic PIRLS (ePIRLS) 2016. Survey-weighted models assessed the extent to which ePIRLS performance was explained by foundational print reading achievement, digital self-efficacy, and engagement indicators. The key finding was that digital reading is best understood primarily as an extension of foundational print-based reading comprehension, not as a wholly distinct construct. The second study used digitalPIRLS2021 data, employing multilevel regression models to investigate how process-related factors, specifically digital self-efficacy and time on task, shape performance. Results showed that achievement was positively associated with socioeconomic resources, home internet access, and time on task, and negatively associated with low digital self-efficacy. Notably, the positive association between time on task and

achievement was stronger among students with low digital self-efficacy. Overall, the results suggest that digital reading is best understood as reading literacy enacted in a technology-mediated environment. For assessment developers, researchers, and policymakers, these findings underscore the need to design and interpret digital assessment systems in ways that distinguish core reading ability from the contextual demands introduced by digital delivery.

14:10:00 - 14:55:00

Transparency in High School Assessment Design and Scoring: A Study of School Base Assessment and the Student Transcript System in Ghana

Location: Trinity Ballroom - Salon 4

Speakers: Maxwell Kwaku Ofosu, Obed Agbenorku

Abstract Transparency in assessment design, criteria, methods, and scoring processes is essential for promoting equity, accountability, and trust in education. An open and student-centred assessment system enhances learning outcomes, increases engagement, and shifts scoring from a closed, teacher-controlled practice to a shared and accountable process. This paper examines the effectiveness of the Student Transcript System for High Schools within Ghana's internal assessment framework, which operates under a national structure of 30% formative and 70% summative assessment. The study explores how the Transcript System strengthens transparency, standardization, and fairness in formative assessment practices. The formative structure incorporates multiple assessment strategies, including individual class assessments (classwork, quizzes, homework), mid-semester assessments, practical/portfolio/performance tasks, individual or group projects (research, case studies, laboratory work, workshops, and presentations), and supervised individual termly assessments. These are administered according to defined national timelines, with scores uploaded to a centralized server to ensure compliance, comparability, and data integrity across schools. A central equity mechanism is the system's verification and correction protocol, whereby scores are printed and publicly displayed for student review before final submission, with corrections permitted only within a structured timeframe. Using policy analysis and system review methodology, the study finds that semester-based digital submissions reduce score forgery, ensure portability of assessment data for transferred students, prevent re-entry manipulation, enforce compliance with prescribed assessment modes, and provide verifiable transcripts for students unable to complete final external examinations. The inclusion of character qualities further promotes holistic accountability. The study concludes that the Student Transcript System advances equity and transparency in Ghana's assessment landscape and offers a replicable model for strengthening integrity in national formative assessment systems.

14:10:00 - 14:55:00

Using AI to Support Effective Use of the Formative Assessment Process

Location: Trinity Ballroom - Salon 3

Speakers: Dr. Edward Roeber

Artificial Intelligence (AI) can support both teaching and learning, especially within the formative assessment or assessment for learning process. This session will show how AI can assist teachers to provide additional support to learners, and how learners can use it to enhance their learning. While formative assessment is a relatively unused assessment practice in the U.S., it is more widely known and used in countries such as Australia, Canada, Great Britain, and New Zealand. Research there has clearly shows the impacts of on-going use of the formative assessment process by skilled teachers with deep understanding of their discipline and pedagogy. Michigan's Formative Assessment for Michigan Educators (FAME) formative assessment process outline was selected because it represents the largest, longest standing formative assessment professional learning program in the U.S. FAME uses this definition of formative assessment: "Formative assessment is a planned, ongoing process used by all students and teachers during learning and teaching to elicit and use evidence of student learning to improve student understanding of intended disciplinary learning outcomes and support students to become more self-directed learners." (CCSSO FAST SCASS, 2017) This session will show several ways in which it can improve learning and instruction. It will be divided into several sections: First, the formative assessment process will be described. Formative assessment is a planned, ongoing process

supporting student learning and teacher instruction, guiding ‘in-the-moment’ instructional and learning adjustments. AI can provide additional support for learning and teaching. Second, a more in-depth description of the five FAME formative assessment process components and 13 elements in the FAME program will be described. AI suggestions for each will be presented. These are viewed as “starters” to get educators thinking about how to use AI. Third, attendees will be encouraged to go beyond these in useful ways and share their prompts and training of the AI engine used to produce desired results. Ideally, such suggested uses will be accompanied with data on how the AI uses positively affected students.

14:55:00 - 15:20:00
Coffee Break with Exhibitors

15:20:00 - 15:50:00
Advancing Assessment Transparency Through Universal Design of Assessment Systems
Location: Trinity Ballroom - Salon 4
Speakers: Dr. Meagan Karvonen
Universal design began as a set of architectural guidelines for developing physical spaces and products that are maximally accessible for all users of the space (Mace, 1997). The Center for Universal Design refined these ideas into seven broad principles that are now included in many regulations and have spread worldwide (Snider & Takeda, 2008; United Nations Convention on the Rights of Persons With Disabilities, 2006). For over two decades, the educational community has attempted to apply universal design concepts to assessment, yet significant gaps remain in current principles and their implementation. While the 2014 Standards for Educational and Psychological Testing identified universal design as a promising approach for improving fairness, existing frameworks like CAST’s Universal Design for Learning (UDL) and the National Center on Educational Outcomes’ Universal Design of Assessment (UDA) have distinct limitations. Furthermore, test developers often claim adherence to universal design without articulating specific procedures or providing evidence of outcomes, resulting in a vague "checklist" mindset. To address these issues, this paper introduces a new framework: Universal Design of Assessment Systems (UDAS). UDAS is a coherent, systems-based approach designed to extend across the full spectrum of educational assessments—from formative classroom practices to large-scale summative and licensure exams. The UDAS framework is built upon three core principles and ten associated guidelines. Principle 1 draws on Self-Determination Theory, asserting that assessments must meet human needs and fulfill the goals of educational systems. Principle 2 focuses on designing assessments for flexibility and learner variability. Principle 3, perhaps the framework's most significant departure from previous models, requires the field to promote equity and take action to mitigate unintended consequences. The presentation will introduce UDAS and provide more in-depth discussion of parts of UDAS that are more distinct from UDA or UDL: (1) maximizing autonomy; (2) broadly defining the construct; and (3) safeguards for fair, responsible and equitable use. This paper primarily addresses the Transparency conference theme through UDAS guidelines that address equity and fairness, data sharing and reporting, and transparency in assessment design. Some of the UDAS guidelines also further aspects of the Technology and Trust themes.

15:20:00 - 15:50:00
From Prompt to Paper: Testing the Validity of AI-Generated Items in High-Stakes Assessment
Location: Trinity Ballroom - Salon 3
Speakers: Lucy Howarth
This study examines how artificial intelligence (AI) can support assessment design by evaluating the quality of items generated by a specialised large language model (LLM) but with no human review. As intelligent systems become more integrated into assessment workflows, it is essential to understand their potential and limitations in producing items that are accurate, appropriately demanding, ethical, responsible and usable in high-stakes contexts. Our aim was to analyse the psychometric behaviour and perceived quality of AI-generated items compared with human-written items, and to consider the implications for using AI to enhance efficiency and accuracy in item development while maintaining quality. A mock GCSE Biology practice test was constructed using paired AI-generated and human-written questions matched on topic, assessment objective and mark tariff. Student responses (n = 203) from four UK secondary schools (serving students aged 11–16) were

analysed using item response theory, alongside facility indices and discrimination indices, to examine thepsychometric properties behind the items. Next, to explore potential human bias and perceptions of quality,teachers – blinded to item origin – rated perceived student ability and attempted to identify whether items werewritten by a human or generated by AI. In parallel, experienced examiners evaluated the quality of each itemand its mark scheme using established assessment quality frameworks. Findings suggested that AI-generateditems performed better than teachers expected, with several items behaving similarly to human-written ones –particularly for 1-mark items. Higher-tariff AI items frequently exhibited poorly calibrated demand, limitingstudent accessibility and undermining measurement accuracy. While teachers were generally unable todistinguish between AI and human items, examiner reviews identified systematic weaknesses in AI-generatedmark schemes, including ambiguity and misalignment with demand, which appeared to drive weakerperformance in several items, suggesting a need for improved prompting. Overall, these findings demonstratethat AI has genuine potential to enhance efficiency in early-stage item development. However, meaningfulhuman oversight remains essential, particularly in calibrating difficulty, refining mark schemes and ensuringalignment with curriculum expectations to maintain validity, fairness, and transparency within teacher-ledintelligent systems. Further research across subjects and item types will help clarify how AI and machinelearning tools can best contribute to accurate, consistent and scalable assessment design.

15:20:00 - 15:50:00

Standardizing Clerkship Grading in Medical Education: A Hybrid Model to ImproveFairness, Comparability, and Validity in Clerkship Assessment

Location: Trinity Ballroom - Salon 2

Speakers: Ebrahim Talaee

Background: Clerkship grading in undergraduate medical education commonly relies on tiered categories (Honors, High Pass, Pass, Fail) derived from multiple components, including clinical performance andstandardized examinations. However, substantial variability in grade distributions across clerkships raisesconcerns about fairness, comparability, and validity. For example, the proportion of students receiving Honorsmay range widely between clerkships, reflecting differences in grading practices rather than true differences instudent performance. Objective: To examine variability in clerkship grading distributions and propose apsychometrically informed reform model that improves standardization while preserving meaningful distinctionsin performance. Methods: We developed a multi-clerkship dataset integrating student-level data across all corerotations, including raw scores (clinical, shelf, composite), tiered grades, clerkship timing (block and order), andstudent characteristics (e.g., longitudinal integrated clerkship vs block-based pathways). Grade distributionswere analyzed across clerkships and components. Additional analyses explored temporal trends, subgroupdifferences, and alignment between subjective (clinical) and objective (shelf) measures. Results: Preliminaryanalyses demonstrate substantial variability in grade distributions across clerkships, particularly in theproportion of Honors grades, suggesting inconsistent application of grading standards. Discrepancies were alsoobserved between raw scores and assigned categories, indicating threshold effects and potential construct-irrelevant variance. These findings highlight limitations in the current criterion-referenced systemwhen implemented independently across clerkships. Proposed Reform Model: We propose a hybrid gradingframework that retains categorical distinctions while introducing a cross-clerkship normalization layer. Thisincludes standardizing scores within clerkships (e.g., z-scores), applying institution-wide thresholds, andimplementing distribution “guardrails” rather than fixed quotas. Centralized grading oversight is recommendedto enhance consistency and defensibility. Conclusions: A hybrid, normalized approach offers a practical anddefensible solution to variability in clerkship grading. By balancing local context with institutionalstandardization, this model supports fairness, improves interpretability, and strengthens the validity of high-stakes outcomes such as the Medical Student Performance Evaluation. This framework is scalable andadaptable to competency-based medical education systems.

15:20:00 - 15:50:00

Technology, Fairness, and Trust in High-Stakes Assessment: Lessons fromAzerbaijan

Location: Trinity Ballroom - Salon 5

Speakers: Vali Huseyn

The growing use of digital technologies in high-stakes assessment has introduced both new efficiencies and

new risks to fairness and trust. This presentation examines a recent case from Azerbaijan, where the use of concealed communication devices during a university entrance examination exposed vulnerabilities in test security and raised concerns about score validity and equal testing conditions. Using a qualitative case study approach supported by international comparisons (e.g., A-levels, SAT, PISA), the study analyses how technologically-enabled misconduct can distort outcomes in centralized selection systems. The findings highlight the importance of integrated quality assurance mechanisms, including pre-exam procedural validation,real-time monitoring, and post-exam statistical analysis (e.g., response pattern and anomaly detection), as well as the growing role of digital testing environments in capturing process data. The study argues that fairness in modern assessment systems is increasingly dependent on the ability to anticipate and manage technological risks, and that maintaining trust requires a systemic approach combining technical safeguards, psychometric evidence, and the promotion of academic integrity.

15:20:00 - 15:50:00

Capitalizing on AI to Build the Human Infrastructure Necessary to Sustain State-Supported School-Based Assessment

Location: Bay

Speakers: Charlie DePascale

“The arc of the history of assessment is moving from institutional purposes and bending toward assessment to inform teaching and learning” (Brookhart and DePascale, 2025, p. 1078).A requisite condition for assessment to inform teaching and learning is for learners and educators to be fully engaged in an assessment process that is integrated into instruction and learning. Such an assessment process requires a fundamental shift from external standardized tests whose primary purpose is accountability to state-supported, school-based assessment programs. Historically, attempts to implement such programs at scale have been unsustainable. Measurement and technological issues have been the primary focus of studies into the unsustainability of these programs. Building the human infrastructure necessary to support and sustain school-based assessment programs, however, has proven to be a more intractable barrier. “Human infrastructure” refers to “the network of people, relationships, knowledge flows, and norms that enable work to actually happen inside an organization” (Digital Wisdom Collective, 2025). With regard to school-based assessment and the theme of this conference, it refers to the time and resources needed to prepare educators, the general public, and other key stakeholders for their use; and manner in which transparency and technology are combined to build assessment systems that learners, educators, and societies can trust. This presentation describes the results of using historical analysis of the root causes that limited the sustainability of assessment programs to identify specific examples of how AI tools can be used to strengthen and support the human infrastructure necessary to sustain such assessment programs. Building human infrastructure is a multi-step and multi-year process which includes: Gaining trust and buy-in from all participants and stakeholders. Allowing adequate time for acclimation and use. Providing continuous support for implementation and use. Providing support for the interpretation and use of assessment results. AI can be most helpful in Stages 3 and 4 by providing direct support to teachers and students at a level and degree that the state will never have the capacity to match. That support begins from initial design and implementation and continues through interpretation and use of results.

15:20:00 - 15:50:00

From Manual to Scalable: Exploring AI-Assisted Item Generation and Validation

Location: Carlton

Speakers: Miguel Prieto

As assessment programs scale across classrooms, institutions, and national systems, the manual creation of high-quality assessment items presents persistent challenges. Traditional authoring processes are time-intensive, costly, and difficult to scale, often resulting in inconsistencies across items and barriers to continuous formative testing. In response, Automated Item Generation (AIG) has gained traction, particularly with the emergence of Large Language Models (LLMs). However, it also raises critical questions around quality, validity, usability, and governance. This session explores how TAO is investigating the responsible use of Generative AI (GAI) to support AIG and validation for both professional assessment experts and everyday educators, with a particular focus on QTI as a unifying model for structure, reuse, and exchange. Rather than treating AI-generated content as free-form output, we examine how curriculum-aligned item templates, prompt

engineering, and structured meta data can guide LLMs to produce assessment items that conform to QTI format and support consistency across platforms. Key challenges in AIG will also be addressed, including the need for high-quality training data, subject-matter expert oversight, and ongoing validation to ensure fairness, reliability,and coverage across diverse learner populations. Developed as part of an open innovation initiative sponsored by the Ministry of Economy in Luxembourg and conducted in collaboration with the Luxembourg Institute of Technology, this work contributes to advancing the broader open-source assessment ecosystem. Attendees will gain practical insights into how AI-driven item generation can be responsibly integrated into QTI-based assessment workflows, balancing scalability with trust, efficiency with quality, and innovation within ter operability.

15:20:00 - 15:50:00

Designing Defensible Performance-Based Assessments for InternationalBenchmarking in Career and Technical Education: Insights from PISA-VET

Location: Bloor

Speakers: Ching-Ni Hsieh

Career and Technical Education (CTE) systems, often known as Technical and Vocational Education and Training (TVET) systems internationally, play a critical role in workforce development, yet many countries lack robust, comparable evidence of how well vocational programs prepare learners for workplace demands. The Programme for International Student Assessment Vocational Education and Training (PISA VET) is a major international initiative designed to benchmark VET performance across countries and identify strengths and gaps in vocational education systems (Iliescu et al., 2025; OECD, 2024; Rausch, 2024). Achieving these goals requires innovative assessment approaches capable of capturing both technical skills and employability skills while maintaining rigor, comparability, and credibility at scale. Partnering with OECD, ETS is developing and evaluating innovative performance based assessment tasks intended for authentic representation of vocational competencies in internationally comparable ways. This presentation reports evidence from a prototyping study of PISA-VET and illustrates how a systematic assessment design process supports the development of defensible and policy relevant VET assessments. Participants included 30 adult students from three automotive VET programs in Portugal. A live demonstration automotive task was developed in close collaboration with VET trainers to ensure alignment with core workplace practices while preserving the standardization required for large scale assessment. Students completed the live demonstration task and then participated in post assessment cognitive interviews designed to elicit systematic feedback on task interpretation, decision making,and perceived task demands. Interview data were translated and coded to examine alignment between intended constructs and observed cognitive processes and to identify task features that introduced ambiguity or unnecessary cognitive burden. Findings indicate that the task largely elicited the targeted competencies while also revealing design features that informed targeted task refinements. The study demonstrates how integrating student feedback through cognitive labs, alongside other design and evaluation activities,strengthens the rigor of PISA-VET assessment and supports credible interpretations for international benchmarking.

15:20:00 - 15:50:00

From Summative Assessment to Formative Feedback? Concept Development in a Re-search–Practice Partnership

Location: King

Speakers: Dr. Joerg Siewert, Kathrin Racherbäumer, Marion Schwehr

FreiDay is a project-based learning format in the context of Education for Sustainable Development (ESD)implemented at the Siegen Laboratory School. Programmatically, it is designed as an assessment-free settingin which formative feedback replaces traditional summative assessment (Rasfeld, 2021; Black & Wiliam, 1998).The school aims to implement this approach, extend feedback-oriented practices beyond FreiDay, and – as alaboratory school – develop transferable models. This paper examines how feedback practices are developed within a research–practice partnership (RPP; Coburn & Penuel, 2016) between educational researchers at the University of Siegen and the Siegen Laboratory School. Building on prior work on the iterative reinterpretation of research findings in collaborative development processes (Siewert & Racherbäumer, in press), the study focuses on how such findings are jointly reinterpreted with teachers for

concept development. The study drawson ethnographic data, including observations and interaction analyses, interpreted using grounded theory. The findings are jointly reflected upon with teachers and serve as a starting point for collaborative development. The analysis reveals a central tension: while the programmatic orientation toward formative feedback implies continuous, assessment-free support of learning, teachers often have limited experi-ence with such practices. As a result, familiar evaluative routines – such as subtle judgments, implicit criteria, or comparative framing –persist even within Frei Day. These practices remain largely unrecognized. Consequently, evaluation continues in classroom interaction without being explicitly marked as such. This leads to a lack of transparency in which criteria and consequences of feedback remain unclear. Such in transparency not only undermines the intended assessment-free envi-ronment – making it a fragile frame of reference for students – but also hinders the development of new feedback practices, as existing routines are perceived as consistent with the approach. This tension forms the starting point for further development of feedback practices within the RPP. The paper presents a typology of implicit evaluative practices as well as a collaboratively developed reflective tool that makes implicit routines visible and accessible for deliberate redesign. In doing so, the paper contributes to the conference theme of transparency by showing how feedback practices can be rendered more explicit and accessible to both teachers and students.

15:20:00 - 15:50:00

A Mixed-Methods Study on the Efficacy of Interactive E-Learning Software on the Academic Performance of Secondary School Students

Location: Trinity Ballroom - Salon 1

Speakers: Dale Peters

This session explores the use of artificial intelligence (AI) and interactive e-learning technologies to enhance assessment design and delivery in secondary education. Drawing on the Farmskillz platform, the presentation demonstrates how AI-driven tools—such as adaptive quizzes, simulations, and gamified tasks—can create more engaging, personalized, and responsive assessment experiences. The session highlights how these innovations address common challenges in traditional assessment, including limited access to resources, time constraints, and varying student learning needs. Emphasis is placed on the role of AI in supporting real-time feedback, improving student engagement, and enabling teachers to monitor progress more effectively. Key insights also consider issues of equity, fairness, and data ethics, alongside the opportunities AI presents for transforming assessment practices. Attendees will gain practical perspectives on integrating intelligent technologies to support more inclusive, flexible, and learner-centered assessment systems.

16:00:00 - 16:30:00

Ensuring Quality of Value-Added Model in Slovenia

Location: Trinity Ballroom - Salon 2

Speakers: Gasper Cankar

Value?added indicators have become an increasingly important tool for evaluating school performancebecause they adjust for many non?school factors that influence student achievement. In Slovenia, thesemeasures have been systematically available since 2014 for upper secondary schools in the general educationtrack, enabling schools to compare students' national assessment results at the end of Grade 9 with theirresults on the General Matura exams. Since 2018, lower secondary schools have also been able to examinevalue?added between national assessments in Grades 6 and 9 across several subjects. It should be noted thatvalueAdded measures in this case are not used for accountability purposes but serve as instruments forinternal self?evaluation and school improvement. Slovenian value?added measures are calculated as theaverage residual between actual and predicted student achievement. The most common prediction techniqueis the "median method," which groups students by their earlier achievement, determines median outcomes foreach group at the later time point, and interpolates predicted values for all students. This method has provenaccessible and easy to communicate to educators and the public. However, emerging evidence suggests thatthe approach may have limitations. One persistent issue is that national average value?added scores tend tobe negative, raising concerns about potential biases or methodological shortcomings. This presentation willunveil research into possible explanations for these consistently negative averages. Three key questionsguided the inquiry: Whether school composition factors such as school size contribute to the patternWhethercontextual characteristics, including municipal socioeconomic

indicators, play a role Whether aspects of the value added computation itself—particularly the median method—may systematically depress results. Given that value added models inherently amplify measurement error by relying on differences between two assessment points, minimizing additional sources of error is essential. To explore these issues, the presentation will employ different regression analyses using five years of value?added data from the National Examinations Centre, combined with municipal data from the Slovenian Statistical Office. All analyses were conducted with strict attention to privacy and legal requirements. Ultimately, the research seeks to clarify the origins of negative national averages and offer guidance for improving value?added methodologies,contributing to broader international discussions on the validity and usefulness of such models.

16:00:00 - 16:30:00

Supporting Appeals in Large-Scale Assessment: A Case Study from Brazil's SAEB

Location: Bay

Speakers: Clara Silva, Raissa Ortega

In large-scale assessment systems where results are formally used in public policy and transfer of financial resources, appeals processes become a critical component of governance. In Brazil, the Sistema de Avaliaçãoda Educação Básica (SAEB), implemented by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, directly informs the Índice de Desenvolvimento da Educação Básica (IDEB), which is linked to national targets established under the Plano Nacional de Educação (Lei nº 13.005/2014). In this context,contestations of results are not merely technical requests, but actions with potential implications for accountability, resource allocation, and policy interpretation. This paper presents an automated model tosupport the submission and review of appeals within SAEB. The model, which is currently under developmentand testing, uses INEP's School Census and SAEB results data to link each appeal to the participant school.The main goal is to support appeal analysis by reviewing submissions and indicating the appropriate responsefor each case. In addition, the model aims to address operational challenges associated with large volumes ofsubmissions, while improving transparency in how appeals are handled. To contextualize this proposal, thepaper also examines two dimensions of SAEB:The scale and complexity of the nationwide paper-basedevaluation system creates significant logistical demands. Although transparency has expanded through publicreporting, microdata dissemination, and the use of IDEB as a communication tool, less attention has beengiven to how systems respond to contestations. By treating appeals as an integral part of assessmentgovernance, this paper highlights the importance of clear procedures and consistency in decision-making. Italso discusses practical considerations related to documentation and data use. The paper contributes toongoing discussions on how to strengthen not only the production and dissemination of assessment results, butalso the processes through which they are questioned and reviewed.

16:00:00 - 16:30:00

From Static Testing to Continuous Measurement: Leveraging AI to Transform Digital Assessment Systems

Location: Trinity Ballroom - Salon 5

Speakers: Stephen Murphy

Educational assessment systems have traditionally relied on sequential, resource-intensive processes—item development, field testing, calibration, administration, and post-hoc psychometric analysis. While this model has supported reliability and validity at scale, it is constrained by cost, time, and limited adaptability in increasingly digital environments. Recent advances in artificial intelligence (AI), particularly transformer- based large language models (LLMs), are enabling a fundamental shift in how assessments are designed, delivered,and maintained. This presentation explores how AI can support a transition from static testing models to a continuous measurement paradigm, where item quality, scoring, and psychometric properties are dynamically monitored and updated. We examine three core areas of innovation. AI in item development enables rapid generation of standards-aligned items, multiple secure variants, and draft scoring rubrics, significantly accelerating development cycles while maintaining human oversight. Emerging approaches to AI-driven difficulty and discrimination modeling demonstrate that LLMs can predict item parameters from text with moderate to strong accuracy, reducing reliance on large-scale field testing and improving test assembly efficiency. AI-enabled psychometric systems support continuous calibration, real-time performance monitoring,and automated analyses (e.g., CTT, IRT, and DIF), enhancing both responsiveness and transparency. The session also addresses key

challenges associated with digital assessment innovation, including validity,fairness, interpret ability, and governance. We emphasize the importance of human-in-the-loop frameworks and Evidence-Centered Design (ECD) to ensure that AI applications align with construct definitions and support appropriate score use. Finally, we consider practical implications for implementation, including scalability for diverse assessment programs, improved accessibility for smaller or emerging systems, and enhanced test security through anomaly detection. This work highlights how integrating AI across the assessment life cycle can create more agile, efficient, and responsive digital assessment systems—while maintaining the rigor and trust required for high-stakes decision-making.

16:00:00 - 16:30:00

ML-Enhanced Cognitive Diagnosis for Transparent Teacher Assessment

Location: Trinity Ballroom - Salon 3

Speakers: Nurseit Baizhanov, Nurym Shora

As digital assessment systems expand, ensuring accurate, robust, and interpretable scoring becomes critical for both decision-making and professional development. This study investigates how machine learning (ML)can enhance cognitive diagnostic modeling (CDM) for teacher assessment, with a focus on improving classification stability and validity (de la Torre & Minchen, 2014). The primary analysis is based on an internally developed mathematics assessment administered to over 3,000 teachers. Domain-level mastery profiles across five content areas were estimated using the DINA and Generalized DINA models. Results revealed heterogeneous mastery patterns, with higher performance in some domains and substantially lower mastery in others. The probability of mastering all domains simultaneously was notably lower than individual domain mastery rates, highlighting the limitations of aggregate scoring and the value of fine-grained diagnostic approaches. To improve model robustness, ML-based methods were applied to refine Q-matrix specifications and enhance classification accuracy. Specifically, L1-regularized logistic regression and decision tree algorithms were used to address potential misspecification, while a Neural Cognitive Diagnosis Model was implemented to capture non-linear relationships among attributes (Wang et al., 2020; Xue & Bradshaw, 2021).These approaches demonstrated improved predictive performance and greater stability of mastery classifications. To strengthen the validity of the findings, diagnostic results from the internal assessment were compared with outcomes from an independent Teacher Knowledge Assessment used within a nation al appraisal/attestation system. Additionally, CDM-based mastery estimates were examined alongside Item Response Theory (IRT) ability scores. Moderate and consistent associations across these measures provide convergent validity evidence, supporting the alignment between ML-enhanced diagnostic classifications and external assessment indicators (Javidanmehr & Anani Sarab, 2017). The findings demonstrate that integrating ML into assessment design and scoring can enhance both the technical quality and inter pretability of results. By combining fine-grained diagnostic modeling with robust validation against independent assessment systems, the proposed approach supports more reliable and transparent use of assessment data.

16:00:00 - 16:30:00

The False Sense of Security: Why More Monitoring Doesn’t Mean More Integrity Location: Trinity Ballroom - Salon 1

Speakers: Brandon Smith, Fraser Reid

In response to growing concerns around AI and misconduct, many assessment systems have expandedmonitoring, increased data collection, and adopted automated detection tools. However, more monitoring doesnot necessarily lead to more trustworthy outcomes. In some cases, it introduces new risks, including increasedfalse positives, inconsistent decision-making, and challenges in explaining and defending results. This sessionchallenges the assumption that more technology and more oversight automatically produce better outcomes. Itexplores how over-reliance on automated detection and excessive data collection can create a false sense ofsecurity while undermining fairness, transparency, and trust. Through applied examples, the session examineshow effective assessment systems are defined not by how much is monitored, but by how decisions are made,validated, and consistently applied. It introduces an alternative approach grounded in proportionality,evidence-based review, and the integration of human judgment. Participants will gain a framework forevaluating current technology use and identifying opportunities to simplify processes while strengtheningintegrity and trust in outcomes.

16:00:00 - 16:30:00

The Relationship Between School-Based Assessment and Final ExaminationScores in Online Schools in South Africa

Location: Trinity Ballroom - Salon 4

Speakers: Dr Busisiwe Mhlongo-Somo

The expansion of online schools has transformed assessment practices in basic education. Traditionally,School? Based Assessment (SBA) has been administered physically under controlled classroom conditions,whereas in online schooling contexts, it is primarily conducted digitally. This alternative assessment modalitywithin the National Senior Certificate (NSC) promotes accessibility and flexibility; however, concerns remainregarding the extent to which SBA scores correlate with final examination scores in online schools. This studyaimed to determine the relationship between SBA scores and final examination scores in online schools. Aquantitative correlational design was employed, utilising data from the Umalusi database drawn from 119 onlineschools from 2023 to 2025 from a single assessment body. About 12 high?enrolment subjects were selected,and analyses were conducted using paired SBA and final examination scores for each year, with Pearson'scorrelation coefficient (r), Spearman's rank?order correlation (r?), and Kendall's Tau (?) computed to examinethe direction, strength and stability of correlations across subjects and schools. Results indicated an overallpositive correlation direction between SBA and final examination scores; however, both the strength andstability of correlations varied over the three years. In 2023, 22.89% of online schools demonstrated strongcorrelations, 59.04% moderate correlations, and 18.07% weak correlations. In 2024, strong correlationsincreased slightly to 25% of online schools, moderate correlations to 66.7%, and weak correlations decreasedto 8.3%. By 2025, a reversal occurred: strong correlations were observed in only 17.53% of online schools,moderate correlations in 18.04%, and weak correlations predominated at 64.43%. Longitudinal analysisrevealed limited stability, with only two schools consistently demonstrating stable strong correlations across thestudy period, while 10 of the 12 subjects exhibited declining, fluctuating, or stable moderate correlation trendsacross most online schools. This study underscores the urgent need to investigate the underlying factorscontributing to weak correlations between SBA and final examination scores in online schools. These inquiriesare critical for identifying systemic, pedagogical, or technological causes that may compromise the equity andfairness in online assessment.

16:00:00 - 16:30:00

From Assessment to Action: Harnessing GenAI to Personalize Feedback and Strengthen Evidence-Based Instruction

Location: Bloor

Speakers: Paul Edelblut

As education systems prepare to "step into the future," the demand for personalized, data-driven insights is growing worldwide. Diverse classrooms require scalable yet individualized solutions when teacher training programs may not meet the demand. This presentation introduces an innovative application of Generative AI to transform student assessments in writing into highly customized reports supporting both classroom teacher sand the students. Traditional assessment reports often lack actionable insights and alignment to evidence-based teaching strategies. With the advent of Generative AI in reporting tools we are able to not only group learners by skills and deficits but also identify hyper-specific and evidence-based strategies to remediate these issues. Connecting the deficits with the best evidence-based remediation, delivered by the classroom instructor, has been the stated goal of assessments for generations and is now possible. Leveraging large language models fine-tuned on rubrics, local educational frameworks and standards, the system evaluates student work against curricular standards, clusters students by proficiency, and produces differentiated teacher reports. These reports include targeted lesson recommendations grounded in research-backed practices linked to the teacher's training to address specific learning needs of each student. These reports upgrade the instruction in each classroom adding value to the assessment, better supporting students and building fidelity with the highest quality teaching pedagogy for each subject. Our preliminary data comes from an Institute of Educational Sciences funded research program out of the University of North Carolina at Chapel Hill. Educators receive standards-aligned diagnostic data and ready-to-use lessons tailored to each studentgroup' sweaknesses connected to their professional development in SRSD (Self Regulated Strategy Development), reducing planning time and improving intervention accuracy and fidelity of implementation. Students receive personalized feedback from the instructor, fostering engagement and ownership of their learning and teaching journey. This approach empowers teachers, whether expert or novice, with precise, research-aligned tools,enabling them to respond effectively to

learning gaps at scale. In the context of the varied educational needs ona global scale, this technology solution offers a blueprint for equitable, high-impact teaching within the local context. By combining AI's analytical power with evidence-based pedagogy, we address the burning question:How can technology truly personalize learning for all?

16:00:00 - 16:30:00
Rethinking Digital Assessment for Infrastructure-Diverse Environments
Location: Carlton
Speakers: Dan Verdick
How can digital assessment operate reliably in environments where bandwidth, devices, and infrastructure are limited? This session explores practical strategies for extending high-quality, open-source digital testing solutions to diverse and resource-constrained contexts. From high-stakes national programs in digitally mature systems to under served or low-connectivity regions, delivering equitable digital assessment requires flexible deployment models grounded in open standards. We will demonstrate how an open assessment platform can operate both centrally and at the network edge — including via lightweight, low-cost hardware such as a Raspberry Pi — enabling assessments to run reliably even where stable internet connectivity, or robust IT infrastructure cannot be assumed. Attendees will see first-hand how this approach reduces infrastructure barriers, expands access, and supports digital inclusion without compromising assessment integrity or data security. The session will also highlight how openness—through transparent governance, shared development,and standards-based architecture —creates long-term sustainability for ministries, education authorities, and EdTech providers. By combining interoperability with practical deployment flexibility, TAO demonstrates how global-scale digital assessment can remain accessible, equitable, and future-ready. Participants will leave with practical insight into implementing open assessment ecosystems that extend digital access even to the most bandwidth-constrained environments.

16:00:00 - 16:30:00
Building Trust Through Culturally Responsive Assessment
Location: King
Speakers: Dianna Ramnath
In contemporary education, there is an overemphasis on performance-based outcomes, through the application of standardized testing and grading, which minimizes the deeper purpose of learning. This paper explores how building trust with culturally responsive assessments can redirect our focus to ensure learner inclusivity and empowering students to become resilient and reflective learners. Culturally responsive assessments and teaching are rooted in constructivist theories of learning, which focus on efforts to participate actively rather than passively. They prioritize student agency, critical inquiry, and culturally responsive engagement that reflects on contextual problems to reframe inquiry-based learning. Examples include experiential learning which immerses students in real-world problem-solving. Socratic seminars which foster connectivity of the subject matter to the real world and consideration for multiple perspectives through discourse. Culturally responsive teaching validates students' lived experiences and identities as assets, if not pillars, in the learning process. These approaches challenge fixed notions of intelligence and ability, encouraging students to embrace challenges, persist through setbacks, and view effort as essential to mastery. This paper proposes that building trust aligned with culturally responsive assessment are instrumental in fostering progress and equity in learning. By shifting the emphasis from extrinsic rewards to intrinsic motivation that is subjective to learners'ability and background, educators can create environments where students value learning over performance. Formative culturally responsive assessment practices, including peer feedback, reflective journals, and portfolio evaluation, concentrate on progress and process over outcomes. The implications are profound: when students recognize the cultural appropriateness of teaching and assessment they are more likely to take intellectual risks, engage deeply with content, and develop the adaptive skills necessary for lifelong learning. For educators, this requires a philosophical and practical shift—embracing flexibility, empathy, and a commitment to equity. Finally, this paper positions cultural responsive assessment not merely as a strategy, but as catalysts for systemic change in education by fostering inclusivity and equity in learning. There is boundless potential to redefine success and unlock the full potential of learners.

17:30:00 - 19:00:00
NCME Reception

Wednesday, September 30, 2026

08:00:00 - 18:15:00
Excursion: Niagara Falls and Niagara-on-the-Lake Experience

Thursday, October 01, 2026

07:00:00 - 16:30:00
Registration

07:00:00 - 08:30:00
Breakfast

08:30:00 - 09:15:00
Measurement at a Crossroads: AI, the Foundations of Assessment, and the Future of Trust
Location: Grand Ballroom
Speakers: Kadriye Ercikan
The integration of artificial intelligence into assessment is challenging the most fundamental aspects of measurement. AI-generated tasks, AI-mediated interaction, and black-box scoring are reshaping all three pillars of assessment as evidentiary reasoning—the construct, observation, and interpretation models—placing validity, fairness, and comparability under unprecedented pressure. These pressures are multiplied in multilingual and multicultural contexts, where algorithms trained on mainstream response patterns may not yield comparable scores across language and cultural groups. Yet AI offers far more than efficiency through automated scoring and content generation. It creates opportunities for better measurement of what matters: interactive, simulation-based tasks that capture response processes, personalization that meets learners where they are, and feedback that places assessment in the service of learning. This keynote argues that these benefits cannot be assumed—they must be demonstrated through rigorous evaluation of efficacy, validity, and fairness. Looking ahead, it outlines a future in which assessment functions as a continuous signal rather than a single event, comparability rests on equivalence of evidence rather than sameness of form, and the field's most urgent work lies in advancing the trust, transparency, and standards that a new measurement paradigm demands.

09:15:00 - 10:00:00
The Elusive Language Component of Skills Assessment: Does AI Help Flag and Resolve Potential Issues?
Location: Trinity Ballroom - Salon 5
Speakers: Øyvind Meistad, Steve Dept
When item authors write assessment questions, they usually write them in their strongest language. They draw on their domain expertise, on their test developer's experience and, importantly, on their language skills. Their writing skills, as well as their socioeconomic and cultural background, contribute to the test form. Idiosyncratic language can add colour, and it can create construct-irrelevant variance: social markers in the language used can introduce a cultural bias, and the level of reading proficiency needed to fully understand the questions may put talented individuals at a disadvantage, particularly if the language of the test is not their heritage language. Traditionally, this has been resolved by resorting to cultural review panels, by having culturally diverse item writers and subject matter experts, and by focus groups. This is costly and time-consuming. In the age of natural language processing (NLP), it is possible to use data from previous cultural review panels, translatability assessments and item health checks to generate an automated screening report of a mature draft of a test or exam. We used a database of known issues and workarounds in assessment items that we compiled and maintained for the past 14 years to program an automated scrutiny application, called TAlly. It took months of testing, identifying false positives, debugging and fine-tuning before TAlly could generate reports that contain entries for all potential issues that resemble previously identified issues. The reports provide a classification-based description of the issue (e.g. potentially ambiguous, unnecessarily complex, double-barrelled question, logical issue) and indicate how it was previously addressed or prevented. This could be through a disambiguation note or by means of a slightly different formulation, without loss of meaning. In this presentation, we demonstrate how the automated item health check was integrated into an assessment authoring platform and how the item writers could process the reports. The case study served as a benchmark to determine the level of human oversight required to make such an automated scrutiny relevant and effective.

09:15:00 - 10:00:00

Mining Insights and Interaction Analysis of Mathematical Problem Solving in Digital Environments

Location: Trinity Ballroom - Salon 4

Speakers: Charles Anifowose, Dr. Robin Sharma

To understand students' mathematical thinking, there is a need to go beyond just correctness (right/wrong answer). It requires assessment approaches that capture the underlying cognitive processes and misconceptions. This article presents a novel framework that integrates APOS theory (Action, Process, Object, and Schema) with large language model (LLM) technology to assess students' mathematical problem-solving in digital environments. The framework enables the automated construction of personalized knowledge graphs that reveal students' conceptual understanding and misconceptions in real time. First, we applied the APOS theory to develop genetic decompositions (GDs) of targeted mathematical concepts. Each GD mapped the cognitive progression associated with the four APOS stages. It defined the thinking processes expected at each level, along with illustrative examples, and observable student behaviors that serve as potential evidence for each stage. Critically, the framework extended the traditional APOS model by incorporating an additional dimension of "misconceptions". We added nodes that represent common conceptual errors associated with each stage. These enriched GDs, complete with detailed descriptions and behavioral indicators, were then used to prompt a large language model deployed within a digital problem-solving environment. Drawing on the theory-informed GDs, the model interprets student behaviors and constructs individualized knowledge graphs. In addition to identifying students' understanding and misconceptions, the knowledge graphs highlight the evidence used to arrive at the results. This process produces a dynamic, theory-grounded visual of each student's mathematical cognition. It provides a snapshot of students' work, which can be used to provide targeted intervention. The proposed framework is topic-agnostic and scalable, applicable across a range of mathematical concepts and problem types. Importantly, it offers teachers a rich, visual representation of each student's mathematical understanding that transcends the limitations of right-or-wrong scoring. By surfacing specific areas of strength and weakness, the model equips educators with actionable insights to design targeted instructional interventions that meaningfully support and advance student learning. To demonstrate its utility, this session will describe four case studies applying the model to items drawn from different mathematical topics in the secondary curriculum. It illustrates how the system surfaces meaningful differences in student understanding and its potential to transform formative assessment in mathematics education.

09:15:00 - 10:00:00

Kazakhstan's Journey in Reforming University Admissions for the 21st Century

Location: Trinity Ballroom - Salon 1

Speakers: Nurseit Baizhanov, Terran Brown

Across higher education systems worldwide, governments are re-examining university admissions models to ensure they remain credible, equitable, and aligned with the skills required in an increasingly complex, digital, and Artificial Intelligence (AI) influenced world. Kazakhstan is undertaking such a reform, embarking on a multi-year journey to modernize its national admissions system while maintaining public trust and transparency in a high-stakes assessment context. This presentation will focus on Kazakhstan's reform pathway. It examines the motivations, design choices, and governance considerations shaping the country's transition from a traditional, subject-based admissions examination toward a future-oriented, skills and knowledge integrated assessment model in an environment that is being shaped by advances in AI. This reform is driven by recognition that existing approaches offered limited insight into higher order thinking, academic readiness, and transferable competencies increasingly valued by higher education institutions and employers. The presentation outlines how Kazakhstan is approaching this transformation through a staged and evidence based process grounded in international best practices. Key elements of the journey include the adoption of evidence-centered design principles, evaluating how the existing assessments compare with relevant international assessments, and early exploration of how Artificial Intelligence may responsibly support assessment development, delivery, and analysis in the future. Central to Kazakhstan's approach is an emphasis on trust and transparency. The presenters will describe how the internal people infrastructure was built to support this reform effort. This is critical for stakeholders to gain confidence that decisions were

evidence-based and defensible, and processes were visible and explainable. The reform process foregrounds clear documentation of design decisions, stakeholder engagement, alignment with international testing standards, and a long-term validity and AI-focused research agenda. These elements are intended to ensure that innovation in admissions assessment strengthens, rather than undermines, public confidence. By sharing Kazakhstan's reform journey, this presentation contributes to international dialogue on how education systems can navigate admissions reform as an evolving process. It offers practical insights for policymakers and assessment organizations seeking to balance innovation with legitimacy, and ambition with caution, as they prepare admissions systems for the future of higher education.

09:15:00 - 10:00:00 | Numeracy Without Borders: Helping Students and Institutions Improve Math Outcomes Through Technology-Driven Assessment for Learning at Scale

Location: Trinity Ballroom - Salon 2

Speakers: Dr. Ovidiu Bagdasar, Tamsyn Murnaghan

Numeracy is more than a set of skills; it is a shared language that opens doors to education, employment, and full participation in society. As institutions confront widening gaps in mathematical readiness, building a community of numeracy, across programs, campuses, and countries, has never been more urgent. This panel explores how cross-institutional and international collaborations can support the development of data-informed numeracy assessment practices that improve learner outcomes. Drawing on the Numeracy Without Borders initiative, Ovidiu Bagdasar (University of Derby) will highlight how coordinated efforts across diverse contexts have been strengthening learner confidence, access, and attitudes. Powered by the award-winning Elevate My Math (EMM) learning programme (featuring diagnostics, interactive remedy lessons and summative assessments), Ovidiu will demonstrate how collaborations between academia and industry can develop novel technologies that engage students, support teachers and improve institutional KPIs. The technology was deployed across the whole institution through the Skills@Derby project and made available in all programmes through seamless VLE integration. Dashboards enabled real-time feedback, while over 2000 digital badges have been awarded for course completion, to learners from all around the world. Furthermore, EMM courses have been embedded in training provided to institutions. Attendees will leave with practical strategies to leverage technology-based assessment at scale, and design frameworks and tools for building inclusive numeracy ecosystems that balance accountability with learner support across diverse educational settings.

09:15:00 - 10:00:00 | Viewing Achievement Data as Pathways Over Time, and Encouraging Tracking of Achievement Issues Back to Source

Location: Trinity Ballroom - Salon 3

Speakers: Maggie Dunlop, Megan Ho

This presentation introduces an approach to using longitudinal datasets derived from public school standardized assessments that encourages a view of student achievement data as pathways over time rather than snapshots indicating points where achievement needs to be fixed. The work was conducted by EQAO, the government agency responsible for developing and administering provincially mandated standardized educational assessments in Ontario public schools. Viewing student achievement as a pathway that develops over time encourages system and instructional improvement planning that covers multiple grades rather than overly focusing on the assessment year in which achievement gaps are observed. The presenters will share findings from a project in which EQAO linked the 2021–2022 and 2024–2025 mathematics assessment outcomes of students who participated in the primary (Grade 3), junior (Grade 6) and Grade 9 EQAO assessments. The Grade 3 to Grade 6 cohort included 115 912 students and the Grade 6 to Grade 9 cohort included 113 033 students. While EQAO has reported cohort outcomes for many years, this dataset was the first set of data to be collected after the COVID-19-related school closures of 2020 and 2021, and thus there has been great interest among educators in Ontario regarding potential learning loss and achievement recovery trends. Common questions have included: Which groups of students are gaining ground, and which students are falling behind? Are trends the same as in the past? In addition to sharing the trends that were observed, the presenters will describe how the data has been shared through research reports, reporting dashboards, and integrated into ongoing outreach and school support work by the agency. For example, presenters will show instructional implications across grades and for some key student groups, and share the

resulting insights that are typically discussed with school boards to inform their future teaching and learning. The presentation will therefore serve two audiences: participants seeking to understand trends in mathematics achievement in Ontario since 2020, and participants seeking an example of how to support educators and education system professionals in viewing student achievement data as a pathway that develops over time.

09:15:00 - 10:00:00 | **Balancing Rigour and Relevance: A Participatory Approach to Assessing Holistic Learning Outcomes**

Location: Carlton

Speakers: Bronwen Magrath, Rachel Outhred

Schools2030 is a ten-year learning improvement programme operating in 1,000 government schools across ten countries. Led by the Aga Khan Foundation - in partnership with governments, school leadership, civil society and research partners - the goal of Schools2030 is to improve student Holistic Learning Outcomes (HLOs) through quality teaching practices. These outcomes encompass cognitive, social, emotional, physical, ethical, and creative competencies that enable learners to thrive in complex and uncertain contexts. A strong and growing evidence base shows that these competencies are not only intrinsically valuable, but also predictive of improved academic achievement, mental health, employability, and prosocial behaviour. This reinforces both the efficiency and equity rationale for embedding HLOs within education systems. Within this vision, Schools2030 supports teachers to assess learners' HLOs and design locally relevant solutions to address barriers to their development. Given the diversity of contexts across Schools2030 countries, the programme adopts a culturally responsive approach to assessment that integrates global evidence with local relevance. Teachers and stakeholders play a central role in defining which outcomes matter, how they manifest in context, and how they can be meaningfully measured. At the same time, global assessment partners provide technical expertise by establishing clear taxonomical boundaries between skills, applying best practices in tool and item development, and ensuring rigour through psychometric analysis. This iterative and participatory process ensures assessments are both contextually grounded and technically robust. A key strength of this approach lies in its impact on teacher engagement and practice. Because teachers co-design and contextualise assessments, they are more likely to use them consistently and effectively. The process deepens teachers' understanding of targeted HLOs, enabling them to recognise and nurture these competencies in learners. Embedding assessment within an iterative innovation cycle further supports reflection, helping teachers interpret evidence, adapt instruction, and continuously improve their practice. By combining stakeholder-driven contextualisation with conceptual clarity and technical rigour, Schools2030 develops assessments that are both meaningful and reliable, supporting more equitable and effective education systems.

09:15:00 - 10:00:00 | **How Assessment Design Shapes Conclusions About Program Effectiveness: Evidence from Zambia's Foundational Skills Program**

Location: Bay

Speakers: Dr. Andreas De Barros

How we measure learning shapes what we conclude about whether programs work. We examine this phenomenon in the context of Zambia's foundational skills program in public primary schools.After two years, the program's effects on targeted foundational mathematics skills are 2.6 times larger than on comprehensive assessments that cover foundational mathematics skills overall, with no detectable transfer to closely adjacent domains. The choice of assessment instrument thus substantially alters the apparent magnitude — and the practical significance — of program impacts. Despite these limited short-run effects on broader measures, exam scores from the Examinations Council of Zambia's Primary School Leaving Examination show positive effects on Grade 7 language and mathematics learning in early adolescence, suggesting that initial assessments may understate a program's longer-run contribution to skills development. The findings raise practical questions for assessment agencies evaluating education interventions: how instrument scope and timing interact to determine what counts as evidence of program success.

10:00:00 - 10:25:00

Coffee Break with Exhibitors

10:25:00 - 11:25:00

Navigating AI's Transformation of Educational Measurement: Perspectives on Responsible Innovation from NCME

Location: Grand Ballroom

Speakers: Christopher Ormerod, Kadriye Ercikan, Okan Bulut, Qiwei He, Susan Lyons

Artificial intelligence is rapidly reshaping educational measurement, presenting unprecedented opportunities and complex challenges for researchers, practitioners, and policymakers worldwide. This facilitated panel brings together leaders from the National Council on Measurement in Education (NCME) to examine how the measurement profession is responding to AI's growing influence across the assessment lifecycle—from item development and scoring to reporting and test security. The panel is organized around four complementary perspectives: NCME President Kadriye Ercikan will present the work of NCME's Presidential Task Force on Responsible AI, which is developing guidance on the ethical, valid, and equitable use of AI in measurement—addressing critical questions of transparency, fairness, and accountability that align with the conference theme. Christopher Ormerod, co-chair of NCME's Special Interest Group on AI in Measurement and Education (AIME) and co-organizer of the 2025 and 2026 AIME conferences, will discuss key research themes emerging at the intersection of AI and measurement science, including machine learning in automated scoring, adaptive testing, and generative AI in content development—areas where innovation must be balanced with rigorous psychometric standards. Okan Bulut, Co-Chair of the 2027 NCME Annual Meeting Program, will describe how NCME's conference programming and training offerings are being designed to build practitioners' capacity for responsible AI use, equipping measurement professionals with technical knowledge and critical frameworks at the intersection of assessment and AI. Qiwei He, NCME's representative on the Joint Committee revising the Standards for Educational and Psychological Testing, will discuss how these foundational standards are being updated to address AI's growing role, ensuring that validity, reliability, and fairness frameworks remain authoritative as AI tools become embedded in practice. Following the moderated discussion, Susan Lyons will synthesize the panel's themes within NCME's broader commitment to supporting measurement professionals through this period of transformation—ensuring the field's workforce has the knowledge, skills, and ethical grounding to navigate an AI-enabled future. The session closes with audience dialogue exploring how international assessment communities can collaborate to develop shared frameworks for responsible AI adoption, ensuring these technologies enhance rather than undermine the trust, transparency, and validity foundational to sound educational assessment.

11:30:00 - 12:00:00

Beyond Minimum Proficiency: Using International Benchmarks in TIMSS, PIRLS, and LaNA to Understand the Continuum of Performance and Inform Policy and Pedagogy

Location: Trinity Ballroom - Salon 1

Speakers: Dr. Ann Kennedy, Dr. Dana Kelly

As part of monitoring education quality and equity under the Sustainable Development Goals (SDGs), countries report the percentages of students who reach "minimum proficiency" in reading and mathematics under SDG Indicator 4.1.1. Adopted by the United Nations, the SDGs have strengthened global expectations for accountability and transparency in education systems, can galvanize political commitment, and track progress toward improved quality and equity. However, reporting only whether students meet a minimum threshold provides limited information for policy and pedagogy. A binary indicator obscures the continuum of student performance and offers insufficient guidance in contexts where many students fall well below minimum proficiency, often low and low-middle income countries, or where substantial proportions exceed it. Policymakers and educators require more nuanced evidence to understand how much students are learning, who is being left behind, and how curriculum and instruction can be strengthened. The International Association for the Evaluation of Educational Achievement (IEA) addresses this need through International Benchmarks of Achievement in TIMSS (since 1995) and PIRLS (since 2001). These benchmarks—Low, Intermediate, High, and Advanced—describe what students know and can do at established points on the achievement scales, reflecting a continuum of performance rather than a single cut score. IEA's newest assessment, LaNA, is psychometrically linked to TIMSS and PIRLS and introduces an additional International Benchmark—Basic—below the existing levels. This benchmark captures foundational skills not previously described, enabling more granular reporting of lower-performing students. The Intermediate Benchmark in

mathematics and the Low Benchmark in reading are used by UNESCO to report minimum proficiency at the end of primary education under SDG Indicator 4.1.1b (TIMSS, PIRLS, and LaNA) and at the end of lower secondary under SDG Indicator 4.1.1c (TIMSS). This presentation demonstrates how benchmark-based reporting provides actionable evidence for policy and classroom practice, illustrated with recent TIMSS, PIRLS, and LaNA results. It will highlight the proportions of students scoring below what is considered “minimum proficiency” for SDG reporting, both overall and by gender, and examine changes in the proportions of students at minimum proficiency over time.

11:30:00 - 12:00:00

Seeing the Whole Picture—Transparency in Assessment Through the Numeracy Project

Location: Trinity Ballroom - Salon 2

Speakers: Kristyn Artibise, Mario Desmarais

Transparency in education and assessment is essential for improving instructional practice, strengthening professional trust, and ensuring that data leads to meaningful action to improve student outcomes. Yet, many educators experience universal screeners as unclear or disconnected from classroom realities. This presentation explores how a system-level approach to transparency—modeled through the Numeracy Project—can support clearer understanding of assessment, more informed decision-making, and targeted interventions that respond to student needs. One persistent challenge in schools is ensuring that teachers fully understand universal screeners: what they measure, how they are designed, and how the results should (and should not) be used. Without shared understanding, screening data can feel imposed rather than informative. Within the Numeracy Project, year round professional development creates structured opportunities for educators to unpack screeners collaboratively, connect results to instructional strategies, and align expectations across classrooms and schools. This sustained learning model moves screening from a compliance task to a meaningful instructional tool. Another transparency gap arises when educators struggle to understand why certain students are not being assessed. Missing data—whether due to absences, exemptions, or system-level decisions—can create confusion and undermine confidence in assessment results. The Numeracy Project emphasizes open discussion of assessment participation, helping teams examine student performance data. By making the “why” behind missing data visible, educators are better equipped to interpret results accurately and plan equitable support. Central to this work is the use of Power BI to explore and share data with school teams. Power BI dashboards transform complex assessment information into accessible visuals that highlight trends, gaps, and opportunities. When analyzed collaboratively with teachers, these data visuals support collective inquiry: teams can ask sharper questions, identify priority learners, and design targeted interventions grounded in evidence rather than assumption. A defining feature of the Numeracy Project is its “teachers following teachers” model, where it extends beyond data into practice itself. By observing colleagues, discussing data together, and reflecting openly on instructional impact, educators build a system-wide culture of learning. This presentation demonstrates how transparency—built through ongoing professional learning, shared data tools, and collaborative practice—creates coherence, trust, and improved outcomes in numeracy education.

11:30:00 - 12:00:00

Digital Assessment Beyond Platforms: Building Low-Bandwidth Systems for Holistic Learner Data at Scale

Location: Trinity Ballroom - Salon 3

Speakers: Crysty Swift, Dr Bridget Mthembu

The global shift toward digital assessment has largely been driven by assumptions of stable connectivity, device access and platform-based delivery. In many education systems, however, these conditions do not exist at scale. This raises a critical question: how can entire assessment ecosystems- not just individual tools- be digitised in low- and no-bandwidth contexts? This presentation draws on the General Education Certificate (GEC) in South Africa as a case study of a transition from pilot to national rollout, supporting over one million learners (in 2026 alone) across a distributed education system, including mainstream, vocational and schools-of-skill. The GEC integrates multiple dimensions of learner data, including academic performance, an innovative mapping of 21st-century competencies and learner inclinations, bringing these together into a holistic learner scorecard intended to support subject choice, career pathways and system-level insight. The

presentation examines how this ecosystem has been designed and implemented through a combination of low-bandwidth digital infrastructure, paper-based tools, and integration with existing national data systems. Central to this is the digitisation of the GEC assessment model, which integrates academic performance, 21st-century competencies and learner inclinations into a holistic framework. The digital platform brings these components together in a structured and accessible way, enabling a seamless rollout across schools and the sector. Drawing on implementation experience from transition to scale, alongside system data, the presentation highlights key lessons from real-world deployment. These include:The importance of designing for actual user behaviour rather than ideal conditions The value of hybrid (paper-digital) models in ensuring inclusion The role of feedback loops in generating actionable insight The complexity of integrating new and non-traditional data types into existing systems The trade-offs between simplicity, usability and data completeness. The findings on GEC demonstrates how low-tech, widely accessible tools can be leveraged to build a coherent, scalable system that captures and connects diverse forms of learner data in ways that support practice rather than compliance, shifting the role of data from static reporting to enabling ongoing professional and system-level conversation so that the learner leaves the education system with a sense of self and their future pathway.

11:30:00 - 12:00:00

Opportunities and Challenges of Digital Learning Journal Data for Reconstructing Students' [Self-Regulated] Learning Processes

Location: Trinity Ballroom - Salon 4

Speakers: Dr. Marlis Pesch, Hannah Bartels

Digital learning environments give rise to new data formats that not only expand empirical access but also help to visualize learning processes and make them usable for educational research and educational practice. Using the example of the University School in Dresden, where self-regulated, project-based learning and digital learning environments are integral to the school's educational concept (Langner & Pesch, 2024; Projektgruppe der Universitätsschule, 2022), this paper demonstrates how systematically implemented digital learning journals generate datasets that enable the representation of learning processes with a process proximity, temporal continuity and a level of data density not previously available (Ritter & Langner, 2026). The paper presents an excerpt from entries in digital learning journals of students in grades 7 to 10 (age 12-16 years), covering a period of nine weeks and approximately 350,000 entries. The data include information on students' learning goals noted at the beginning of each learning session as well as their reflections at the end and builds on the idea to analyze processes of self-regulated learning by diary data (Schmitz & Wiese, 2006). The opportunity of structuring, aggregating and linking large, process-generated datasets with other quantitative and/or qualitative data sources (e.g. self-created timetables , goal agreements with learning guides or competency feedback) alters the logic of knowledge production (Vogl, 2023). The data presented visualize how students structure, monitor and adapt their learning processes at different levels. As a result, self-regulated learning becomes visible through data-driven reconstructions of processes (Cleary et al., 2012), rather than through simple learning outcomes. Besides showing central findings, the concern of this paper is to exemplify how digital large-scale datasets establish new forms of knowledge generation, thereby transforming both scientific analysis and educational practice and contributing to the emergence of new knowledge. This knowledge can then be used to adjust the support and guidance to promote self-regulated learning of students in everyday school life. The paper also addresses and discusses methodological and research-ethical challenges arising from working with longitudinal, high-dimensional, and personally identifiable datasets.

11:30:00 - 12:00:00

Promoting Fairness for Assessment of English Writing with Small Language Models

Location: Trinity Ballroom - Salon 5

Speakers: Dr. Shafiza Mohamed, Paul Edelblut

Global advances in natural language processing (NLP) have been largely driven by the development of large language models (LLMs) trained on massive datasets. However, these datasets overwhelmingly reflect dominant Western, white, and male perspectives, primarily from American and European contexts. As a result, such models often fail to accurately interpret, score, or generate responses that reflect the nuanced realities of local languages, cultures, and communicative norms. This misalignment is particularly problematic in the local

Asian contexts of English assessment where local linguistic elements correctly incorporate and are culturally appropriate. This study argues for the importance of local context scoring of written English responses especially given the very large data for evaluating texts not by globalized and biased benchmarks, but by the standards, priorities, and idiomatic expressions that matter in specific communities. Using examples from Malaysia and New Zealand how small language models (SLMs), trained or fine-tuned on regionally grounded data, could explicitly capture local semantics, discourse markers, and culturally specific reasoning structures. Unlike LLMs, which inherit the blind spots of their scale and training data, SLMs offer flexibility, transparency, and adaptability, enabling stakeholders to embed fairness and inclusivity into their assessment pipelines. By centering local context scoring and advocating for small-scale, community-informed approaches, we propose a shift away from the “bigger is better” paradigm toward one where language technologies are aligned and embraced with diverse human experiences.

11:30:00 - 12:00:00

Mapping SDG 4.1.1 Minimum Proficiency Level Cut Points onto Local Assessment Scales Using Paired Comparisons: A Comparison of Scale Linking Transformations

Location: Bay

Speakers: Nathan Zoanetti

Background: In the context of the United Nations’ Sustainable Development Goal (SDG) 4 on quality education for all by 2030, monitoring countries’ progress is important to ensure that the goals are eventually achieved. Data from national large-scale assessments are often central to this monitoring. However, reporting against SDG 4.1.1 requires defensible methods for linking local assessment scales to global Minimum Proficiency Level (MPL) benchmarks. The Pairwise Comparisons Method (PCM) Toolkit supports this process by placing items from established scales on which the MPL benchmarks have already been located and items from local assessments onto a common latent scale. However, limited guidance exists on the choice of scale linking transformations that mitigate bias in the cut-score mapping in practical applications of the method. Method: This study evaluates alternative scale linking transformations within the PCM Toolkit framework, with particular emphasis on comparing mean-mean, mean-sigma and chained linking of MPL cutpoints from LPS scales onto local assessment scales. In the first stage, a common pairwise scale is constructed using the Bradley–Terry–Luce model to estimate item locations, scale reliability, and diagnostic statistics for items and judges. This pairwise scale provides an empirical anchor for subsequent linking. In the second stage, different linking transformations are applied to transfer MPL cutpoints. A simulation study is used to examine how differences in scale dispersion arising from the two data generating methods are handled by different linking transformations. A de-identified case study using data from a national assessment of reading and mathematics in one African country is also included. Results: Results indicate that mean-sigma transformations offer advantages over mean-mean equating when linking pairwise and response-based scales, when scale dispersions differ, as they have typically been observed to do. Conclusions: The findings support moving beyond mean-mean linking within the PCM Toolkit and highlight critical PCM study design considerations for minimising bias. Practical recommendations are provided to support transparent, reproducible, and statistically defensible SDG 4.1.1 reporting using PCM-based linking. The conclusions from this study have wider relevance to other education policy contexts where jurisdictional alignment and benchmarking are necessary and where traditional equating designs are not possible.

11:30:00 - 12:00:00

Monitoring of Student Progress in Reading: Case of Short-Term Training for BINOM (School Network) Language Teachers

Location: King

Speakers: Bilimgul Bayekeshova

As it is known that the main aim of ILSA is to develop an empirically based foundation of knowledge for improvement of educational policy and practice. Pre-analysis of the ILSA (PIRLS, PISA, PIAAC) results have shown moderate achievements in reading quality of primary school students, reading competence of 15-year-old young people and the quality of the reading component of the adult population in Kazakhstan. This research explores how far L1 teachers are familiar with the rigorous methodology and effective reading literacy assessment instruments of International large-scale assessment (PISA). Reading literacy has been recognized as a ‘human value’ (Guerra, 2013) that brings progress, awareness and development especially, now when

there is an increase in the jobs for non-routine analytical and interpersonal skills (weforum.org. 2015). Notwithstanding the existing dilemmas, remaining criticism and limitations as expressed in the international handbook, it is evident that «tests and exams provide huge benefits to organizations, to individuals and to society» (Kleeman). Among plethora of potential benefits are the capacity building in item development, fair assessment practices and educational accountability and policymaking. This paper seeks to analyze the knowledge and understanding of language teachers about PISA, the world's most important test in changing education globally and how it affects the assessment of individual student ability in reading and an education system as a whole. As emphasized by Wolff «participating in international exams is not financially prohibitive and can offer value, as long as the results are enforced to reform the curriculum and to develop teacher training programs» (Napier, 2014). Considering the assessment framework guidelines, benchmarks and measurements of reading achievement in PISA, small-scale survey instruments: available in Kazakh PISA reading literacy, global competence, creative thinking tests (https://www.oecd.org/en/about/programmes/pisa/pisa-test) have been utilized with follow up training questionnaires. Besides, classification and components of assessment tasks are analyzed to inform the performance of low and high order reading skills. The findings help us describe the current state of school curriculum as well as the progress in professional development of in-service language teachers with respect to helping students achieve higher reading literacy levels.

12:00:00 - 13:00:00
Lunch

13:00:00 - 13:30:00
Building Trustworthy, Transparent AI in K–12 Assessment: A Practical Roadmap
Location: Grand Ballroom
Speakers: Rochelle Michel
As AI enters K–12 assessment, leaders must balance innovation with public trust. This presentation delivers a practical, stakeholder-centered roadmap to ensure trustworthy, transparent AI adoption across state and local assessment systems. Drawing on proven implementation practices (e.g., establishing achievement cuts, statewide growth measures, automated essay scoring), we define roles for state boards, assessment staff, teachers, parents, and service providers and clarify lines of authority for approval and accountability. Central to the approach is transparency: a repeatable decision heuristic (risk × impact 2x2), operational definitions and examples of risk and benefit, and a phased rollout with explicit success markers and safety nets. We emphasize communicating the "why" through simple logic models ("if this, then that") and running buy-in campaigns focused on process integrity as much as technical solutions. Attendees will receive concrete tools including: A stakeholder role map. Decision-support criteria for prioritizing low-risk "easy wins" versus higher-risk pilots. Approval pathway checklists. By consistently returning to agreed principles and the heuristic, assessment leaders can deploy AI that advances learning while preserving trust and ensuring transparent, accountable decision-making. Ideal for policymakers, assessment leaders, and district implementers seeking a principles-driven path to responsible AI in assessment.

13:30:00 - 14:00:00
From Assessment to Accountability: Unifying Teams and Accelerating Insights Across the Assessment Lifecycle
Location: Grand Ballroom
Speakers: Charles Anifowose, Dr. John Ladan
The ultimate goal of educational assessment is to transform student performance data into meaningful, actionable support for learning. Aligned with the IAEA 2026 theme of advancing trust, transparency, and technology, this session explores how seamlessly connecting every stage of the assessment lifecycle unlocks system-wide efficiency and innovation. In this session, Dr. John Ladan and Charles Anifowose demonstrate how modernizing data pipelines bridges the gap between assessment and accountability. Rather than treating data management as a post-test administrative task, this presentation showcases how an AI-powered Data Lakehouse architecture breaks down operational silos and empowers every team across the assessment lifecycle simultaneously. By enabling continuous, secure data flow, teams involved in item authoring,

registration, test administration, scoring and marking, data analysis, and reporting gain immediate, synchronized access to live insights. Operational logistics become more responsive, item banks are continuously refined using real-time performance metrics, and computation pipelines drastically reduce time-to-report from months to hours. Through a demonstration of a unified, real-time data environment, attendees will gain practical insights into how cross-functional collaboration and advanced data architecture transform the assessment lifecycle to elevate outcomes for all learners.

14:10:00 - 14:55:00

Assessment Data in Action: Integrating Reporting, Research and Engagement at Scale

Location: Trinity Ballroom - Salon 5

Speakers: Dr. Jennifer Hove, Joanne Rinella

Across education systems, large-scale assessments are increasingly expected to inform improvement planning—not simply report results. This presentation examines how Ontario's Education Quality and Accountability Office (EQAO) has designed an integrated assessment data and research ecosystem that supports instructional decision-making, system planning and sustained educational impact. Grounded in EQAO's recent work, the session highlights three interconnected pillars: modernized reporting of assessment results, applied research that makes learning visible and purposeful knowledge mobilization through outreach and partnership. Together, these elements illustrate how assessment agencies can translate data into action at scale: Modernized Reporting to Extend the Value of Assessment Data EQAO's digital, interactive reporting moves beyond static score release to support improvement-focused use of assessment results. This session will highlight how EQAO's dashboards integrate assessment data into school and board planning, respond to educator questions through flexible filters and drill down views, and extend available insights by reporting sub-scores alongside cohort results over time. Together, these features position reporting as an active decision support tool embedded within continuous improvement processes. Applied Research That Makes Learning Visible EQAO's applied research program leverages large-scale assessment data to deepen understanding of student learning. Examples include mathematics item-level distractor analyses that reveal recurring misconceptions across grades, longitudinal pathway studies that trace achievement trajectories and system-level research identifying practices associated with sustained improvement. The focus is on research explicitly designed to inform instruction, intervention and policy—not solely describe outcomes. Knowledge Mobilization Through Outreach and Partnership EQAO embeds research and reporting insights within a deliberate knowledge mobilization strategy that emphasizes collaboration, relevance and trust. This includes professional learning sessions, targeted board partnerships and curriculum-linked resources. A key feature is collaboration with school board researchers through seconded positions on the EQAO research team, strengthening practical relevance, shared analytic capacity and connections between provincial research and local improvement. By engaging educators, leaders and researchers as partners, EQAO helps ensure assessment insights are understood, trusted and applied across classrooms and systems. The session offers a transferable model for assessment organizations seeking to design ecosystems where data, research and engagement work together to produce measurable educational impact.

14:10:00 - 14:55:00

Delivering a Fixed-Form English Language Proficiency Assessment to English Learner Students Across the Hearing Spectrum

Location: Trinity Ballroom - Salon 3

Speakers: Michelle McCoy

Administering standardized English language proficiency assessments to students across the hearing spectrum presents a persistent challenge in large-scale assessment systems. Many Deaf and Hard-of-Hearing English learners rely on signed language as their primary mode of communication, while English proficiency assessments are designed to measure proficiency in spoken and written English. This creates a fundamental tension between equitable access and construct validity, particularly in fixed-form digital testing environments. This paper presents a practice-based analysis of test administration decisions across the four domains of language–listening, reading, writing, and speaking–within a large-scale, fixed-form assessment system currently in use in the United States. Drawing on guidance from research on assessment access for students across the hearing spectrum, this paper synthesizes accessibility frameworks, administration policies,

and operational constraints to examine how test administrators can support access without compromising score interpretation. The analysis is grounded in a central construct-based distinction between test directions and test content. Test directions describe procedural aspects of the assessment and may be delivered in signed language to ensure examinee understanding. Test content represents the construct being measured and may not be signed or otherwise translated without altering validity. This distinction is applied across all four domains, with domain-specific scenarios illustrating permissible practices, common misunderstandings, and persistent areas of tension. This work reflects sustained experience in large-scale assessment design, where sensitive access decisions for Deaf and Hard-of-Hearing learners remain part of a recurring and unresolved operational challenge despite ongoing refinements in accessibility policy and digital delivery systems. Implications for assessment leaders and practitioners include the need for clear pre-administration planning, explicit documentation of accommodations, and consistent application of construct-based decision
rules. Sustaining trust in digital assessment requires transparency not only in what supports are allowed but also why those decisions preserve the integrity of the construct being measured. This work aligns with the conference theme of transparency by emphasizing explicit, construct-based decision rules that distinguish between permissible access supports and those that alter score interpretations.

14:10:00 - 14:55:00

Re-examining South Africa's grade 12 mathematics performance by question from 2014-2025: Is learner performance deteriorating over time?

Location: Bay

Speakers: Biki Lepota

High-stakes examinations, such as South Africa's exit-level examinations for school leavers, demand fairness in every respect. This study examines performance in the National Senior Certificate (NSC) Mathematics examinations by question for the period spanning 2014-2025. Because the Department of Basic Education (DBE) accounts for approximately 95% of the NSC examinees, the DBE's Mathematics examinations are a particular case in point. That is to detect any changes in question difficulty across years, which might treat examinees unfairly and thereby suggest a problem in using the marks achieved to draw comparisons from one year to the other. The paper is informed by the question: "What does the South African grade 12 Mathematics performance by question tell us about the changes or stability in question difficulty during the period 2014-2025?" To answer the question, the study analysed:The DBE's NSC Technical and Diagnostic Reports for Mathematics covering the period 2014-2025 and The data from Umalusi research project broadly titled "Maintaining standards" conducted over the same period. The analysis yielded useful insights. To start with, enrolments averaged approximately 250 000 over the period, with a downward trend beginning in 2022. Second, the allocation of marks to different topics closely aligns with policy. Concerning cognitive demand, the exams have a greater proportion of questions testing routine procedures and a smaller proportion of knowledge, complex procedures and problem-solving questions than the CAPS' stipulation. The question arising is: Why is there a drop in overall performance in 2025. The evaluation of difficulty levels revealed that the exams have preponderance of questions of moderate difficulty. The papers are getting easier since 2022. Fourthly, the failure rate in Mathematic has revolved around 46% since 2014 while the distinction rate averaged 2.8%. The failure rate decreased to 36.6% in 2023 and decreased further in 2024 to 30.9% before rising to 36%in 2025. Correspondingly, the distinctions rate rose to 3.4% and 3.9% in 2023 and 2024 respectively before declining by 2% to 1.9% in 2025. Concerning performance by question, the results varied widely from
year-to-year, thereby bringing into question the reliability of the examinations. The results can be used by assessment bodies to improve their examination setting processes while Umalusi can use them to enhance its quality assurance processes. Keywords: mathematics, examinations, grade 12, performance, NSC

14:10:00 - 14:55:00

One Platform, Global Partnership: Delivering PISA 2025 Online and Offline

Location: Trinity Ballroom - Salon 2

Speakers: Miguel Prieto

PISA 2025 marked a major milestone in international large-scale assessment delivery as the first cycle implemented on a fully digital assessment platform, while still supporting offline testing where required. With participation across 90+ countries and delivery in 54 languages, this cycle represented the most complex implementation in the history of PISA. Addressing this complexity required close collaboration between ACER

and Open Assessment Technologies (OAT), combining ACER's expertise in large-scale assessment design and international implementation with OAT's technical leadership in the development and scaling of the TAO digital assessment platform. Together, the partnership delivered a centralised assessment solution capable of operating at unprecedented scale while supporting diverse regional contexts, innovative assessment types, and the full rigour of OECD technical standards. This session will present the technical, operational, and organisational complexities encountered when deploying and scaling a high-stakes digital assessment internationally. We will explore how implementation challenges were identified and addressed collaboratively, including adaptations to established national workflows and the introduction of assessment components such as listening, speaking, and highly interactive digital tasks. The session will examine real-world constraints that shaped the programme, including variable infrastructure reliability, regional data requirements, hosting considerations, and the need for continuous operational support across global time zones. We will also discuss how these factors influenced architectural decisions, delivery models, governance structures, and operational workflows, and how risks were mitigated through coordinated field trials, proactive readiness checks, controlled configuration management, and robust monitoring. A central theme is the robustness and scalability achieved through open standards and shared platform governance. Attendees will learn how platform interoperability and joint operational oversight enabled a shift from locally managed workflows to a single, standardised delivery model — reducing operational burden and data-collection risk. We will discuss how this architecture was validated at scale, supporting concurrent peaks of approximately 20,000 users and close to 900,000 assessment and survey sessions while ensuring secure, auditable data capture. Attendees will gain practical insights into designing, governing, and scaling standards-based digital assessment platforms through effective cross-organisational collaboration under demanding international conditions.

14:10:00 - 14:55:00

Reframing Uncertainty as a Design Space: Exploring the Futures of Wearable Technologies in Learning and Assessment

Location: Carlton

Speakers: Dr. Fawaz Abu Sitta

As wearable technologies mature and converge with artificial intelligence, digital twins, and augmented realities, education is entering a period of profound uncertainty that cannot be addressed through forecasting or incremental planning alone. This paper presents a foresight research study that deliberately reframes uncertainty not as a gap to be reduced, but as a strategic design space to be explored, inhabited, and shaped. The study positions future thinking as a translational practice, capable of informing policymakers and educators about long-term implications of emerging human–machine ecosystems. The research uses a participatory foresight approach to explore uncertainty. Rather than predicting specific technologies, it combines horizon scanning, participatory workshops, and structured synthesis to identify plausible futures for learning and assessment. In workshop settings, participants; including students, teachers, policymakers, and technology experts, are engaged in activities that explore uncertainty through multiple lenses: realist, dreamer, and critic. This approach enables tensions, ethical dilemmas, and institutional constraints to be articulated alongside imaginative possibilities. The aim is not technological advocacy, but the development of anticipatory capacity and the responsible exploration of futures beyond wearables and augmented humans. The study captures a set of emerging signals across visual, auditory, tactile, olfactory, biometric, movement-based, and sense-making domains. These signals include:AI-powered AR glasses that adapt visual content based on cognitive load; Eye-tracking used as a proxy for attention and comprehension; Translation earbuds and sound-filtering wearables reshaping multilingual classrooms; Haptic feedback systems guiding motor learning and emotional regulation; Olfactory interfaces supporting immersive and vocational learning; movement tracking that assesses learning processes rather than outputs; andBiometric systems that detect cognitive overload and inform adaptive assessment timing. Emerging signals suggest that wearable technologies may act as cognitive mirrors, allowing learners to interpret their own data through dashboards of their learning, while learning and assessment systems adapt continuously to the learner's evolving state. The paper argues that these signals indicate a shift from screen-based, standardised learning and assessment toward embodied, adaptive, and immersive learning ecosystems. However, they also surface significant ethical, institutional, and governance challenges, particularly around surveillance, agency, and the role of assessment. By filling the space of uncertainty with insight rather than certainty, this research contributes a futures-informed framework for navigating the plausible implications of wearable technologies for learning and assessment over the next 20 years.

14:10:00 - 14:55:00

Beyond the Multiple-Choice Question: Implementing Multi-Modal Assessments in Digital Environments; Opportunities, Obstacles, and Operational Realities

Location: Trinity Ballroom - Salon 4

Speakers: Sanketa Anand

While multi-modal assessments — combining text, simulation, audio, video, and performance tasks — offer a significantly richer measure of competence than traditional MCQ formats, the gap between their promise and operational reality remains large. This abstract is structured around three honest angles: The opportunity: Multi-modal formats improve construct validity and better assess higher-order skills, with evidence from real deployment programs across professional certification and national exams. The obstacles: Bandwidth constraints that exclude candidates in low-resource environments, higher development costs, AI scoring limitations, and the psychometric validation burden that high-stakes programs must meet before adopting new item types. The path forward: An incremental, evidence-driven adoption framework: start with lower-stakes components, build validity evidence progressively, and use hybrid offline-capable delivery models to preserve access. The central message is that multi-modal assessment is not a binary leap. Programs that treat it as a gradual, validity-led evolution can improve what they measure without creating new inequities for candidates in infrastructure-poor environments.

14:10:00 - 14:55:00

Leading System-Level Assessment Reform Through Trust-Based Practice

Location: Trinity Ballroom - Salon 1

Speakers: Adelee Penner, Katherine Mann, Sharon Friesen

Educational systems face increasing pressure to demonstrate accountability while also strengthening assessment practices that support learning, professional judgement, and equitable student outcomes. Leading System-Level Assessment Reform Through Trust-Based Practice explores how one educational system approached this tension by positioning trust—not compliance—as a core condition for large-scale assessment reform. Drawing on a multi-year initiative to develop and implement an outcomes-based assessment strategy, the session examines how governance, professional learning, communication, and implementation were aligned to create greater coherence while preserving space for professional judgement. Scenario-based sensemaking surfaced assumptions, tested interpretations of policy and practice, and developed shared language and expectations before system-wide implementation. The session highlights how transparentdecision-making, intentional pacing, collaborative calibration, and the use of assessment evidence for professional learning helped strengthen trust. It also considers ongoing tensions related to educator readiness, stakeholder expectations, consistency, and professional judgement. Participants will leave with practical insights into how system leaders can create the guardrails and relational conditions needed to support assessment reform that maintains public confidence while advancing learning and continuous improvement.

14:55:00 - 15:20:00

Coffee Break with Exhibitors

15:20:00 - 15:50:00

Beyond the Process Illusion: Co-Constructing Transparency in Assessment Design

Location: Trinity Ballroom - Salon 2

Speakers: Mélanie Renaud

Assessment frameworks worldwide encourage teachers to gather evidence not only of what students produce but of how they think and learn. Yet in practice, most evidence categorized as "process" (drafts, reflections, oral defenses) consists of artefacts collected at predetermined checkpoints. What some call process assessment is more accurately described as longitudinal product assessment. The thinking, the decision-making, the metacognitive negotiation behind those artefacts remains largely invisible, not because it is inaccessible, but because the methods used to capture it rarely go beyond the artefacts themselves. Policy documents such as Ontario's Growing Success (2010) invoke the process/product distinction without defining it with precision,

leaving teachers to construct a binary such frameworks never intended. This presentation examines how transparency in assessment depends less on the volume of trace data than on the explicit clarification of what counts as evidence, how it will be interpreted, and how students participate in making their learning visible. It is situated within an ongoing K–12 action-research project on AI integration in a French-language school board in Ontario (CEPEO) and informed by pre- and post-intervention Likert-scale surveys for teachers and students, classroom observations, participant discussions, and assessment samples. It explores assessment practices that incorporate traceability journals, metacognitive questioning, student justification of decisions, and flexibility in how learning is collected and demonstrated.Emerging insights from this work suggest that these practices help reframe triangulation not simply as a technical requirement, but as an interpretive act supported by conversations, observations, and products. From this perspective, the teacher's role shifts from assessing multiple products to facilitating evidence generation, prompting metacognitive explanation, and coaching students in the articulation of their decisions and thinking – practices that are simultaneously teaching and assessment. The presentation argues that transparent assessment is not merely the visibility of more data, but the shared, explicit co-construction, in real time, through questioning, dialogue, and metacognitive exchange, of what constitutes meaningful evidence and how that evidence is interpreted in relation to learning goals and evaluation criteria. This shared understanding is what makes classroom assessment more open and credible.

15:20:00 - 15:50:00

Advancing Assessment for Young Learners in a Digital Connected World

Location: Trinity Ballroom - Salon 5

Speakers: Xiangdong Wei

The Hong Kong Examinations and Assessment Authority (HKEAA) has administered the Hong Kong Diploma of Secondary Education Examination (HKDSE) since 2012, offering globally recognised academic qualifications and diverse pathways for secondary school graduates. Providing a broad spectrum of academic, applied learning, and other language subjects, the HKDSE promotes holistic development, digital education and cultivates biliterate and trilingual talents, reinforcing Hong Kong's role as a bridge connecting with the world. This presentation will delve into the latest strategies and best practices adopted by the HKEAA to enhance the quality of public examinations and assessments. Key initiatives include School-based Assessment and the Diagnostic Feedback System (DFS), which support comprehensive learning outcomes and foster collaboration with school partners. This session will also highlight initiatives to strengthen the international recognition of the HKDSE, preparing students to thrive in higher education worldwide and contribute to global communities.

15:20:00 - 15:50:00

Using Diagnostic Classification Models to Improve Instructional Decision-Making

Location: Trinity Ballroom - Salon 3

Speakers: W. Jake Thompson

Educational stakeholders increasingly demand assessments that do more than rank students along a single score continuum. They want actionable, skill-level feedback that teachers can use immediately to improve instruction, while also reducing the student testing burden. Diagnostic classification models (DCMs) are a compelling psychometric solution to this challenge. DCMs are psychometric models that provide fine-grained, skill-level mastery information. Additionally, because DCMs provide categorical results (e.g., mastery vs. nonmastery), they can produce highly reliable classifications from short assessments that can be embedded into normal instructional cycles. Despite these benefits, the operational adoption of DCMs remains limited. This presentation describes how DCMs can be leveraged to support better instructional decision-making across multiple stakeholder groups. We begin with a description of DCMs and what makes these models uniquely suited to the task of improving instructional decision-making with assessment data. Unlike item response theory, which places students on a continuous ability scale, DCMs yield a binary mastery profile for each measured skill—providing fine-grained, actionable feedback rather than a single aggregate score. We discuss key modeling decisions practitioners must navigate, including the choice between compensatory and noncompensatory model specifications and the implications of attribute hierarchies for both model estimation and score reporting. We then move to the central focus of the presentation, which is results reporting. We describe and illustrate multiple reporting formats designed for different stakeholder groups, from individual student learning maps that visualize skill progressions to classroom-level roster views that support differentiated instructional grouping. We'll share applied examples of DCMs being used in innovative

through-year assessment models such as the Pathways for Instructionally Embedded Assessment (PIE) project and the Dynamic Learning Maps (DLM) alternate assessment system. These programs demonstrate how DCMs can be used to refocus the purpose of assessment as being in the service of student learning. Attendees will leave with a conceptual foundation for DCMs, concrete examples of diagnostic reporting in operational contexts, and a realistic picture of both the promise and the challenges of implementing DCM-based assessments at scale.

15:20:00 - 15:50:00

Designing for Defensibility: What Makes Assessment Outcomes Stand Up to Scrutiny

Location: Trinity Ballroom - Salon 1

Speakers: Brandon Smith, Fraser Reid

As assessment systems become increasingly digital and distributed, the concept of defensibility has emerged as a critical foundation for trust. This session examines what makes an assessment outcome defensible in practice, particularly in environments where identity verification, monitoring, and decision-making span multiple systems. Drawing on applied experience across credentialing and high-stakes assessment programs, the session explores how decisions are made, documented, and supported when outcomes are challenged through audit, appeal, or review processes. A lifecycle-based framework will be introduced to examine how trust is established before, during, and after the assessment event. Key areas include:How evidence is collected; How decisions are reviewed; and How consistency is maintained across cases. Common failure points will also be explored, including over-reliance on automated signals, inconsistent adjudication, and insufficient documentation. The session emphasizes how system design, human judgment, and process alignment contribute to outcomes that can be clearly explained and defended. Participants will gain insight into how assessment systems can be designed not only to detect irregularities, but to support decisions that withstand scrutiny and reinforce confidence in results.

15:20:00 - 15:50:00

Inclusive-by-Design: Reimagining Digital Assessment Ecosystems in South African Higher Education

Location: Bay

Speakers: Nothando Ntshayintshayi

Are our digital assessments truly inclusive, or are they just digitized versions of the status quo? Assessment is more than a test; it is an ecosystem shaped by the interplay of technology, pedagogy, and governance. In South Africa’s higher education landscape, the shift to digital environments presents a paradox: digital tools offer unprecedented opportunities for innovation, yet they frequently fail to address the complexities of student marginalization. In this session, we investigate how to design digital assessment ecosystems that prioritize multidimensional social inclusion. By synthesizing current research through a five-dimension framework, covering inclusive design, progression, recognition, outcomes, and participation, we identify the critical factors that separate successful digital transformation from exclusionary testing. We move past the basics of access and delve into the necessary governance principles that safeguard student privacy and academic integrity. Attendees will leave with actionable guidelines to transition from traditional digitized testing to a coherent, pedagogical assessment approach that uses technology to provide rapid, helpful feedback and diverse pathways for student success. Let's explore how to build "inclusive-by-design"; assessment practices that empower the student.

15:20:00 - 15:50:00

Data Sovereignty and Algorithmic Justice in Technology-Enabled Assessment: A Context-Responsive Privacy Framework for Southeast Nigeria

Location: Carlton

Speakers: Dr. Kelechi Ihechu, Patience Agommuoh, Prof. Ursula Akanwa

The widespread implementation of technology-enabled assessment (TEA) systems in educational institutions in Southeast Nigeria opens enormous possibilities for scalability and adaptive education but also creates significant challenges for data safety and ethical governance. This research analyzes the discrepancies between internationally accepted standards of data privacy and the reality of TEA use in low-resource settings, particularly in Southeast Nigeria. This analysis is based on the findings of a survey of 43 research articles from reputed academic journals, international standards, and legal regulations/regulatory instruments. The

researchers proposed a privacy protection framework that is based on data sovereignty principles and algorithmic justice. The framework suggested technical solutions, such as federated learning systems and differential privacy methods, paired with governance structures suited to the region's specific challenges and needs. As a result of the research, three major problems were identified: (1) there are no sectoral ethical frameworks for the use of educational technologies; (2) auditing algorithms and detecting bias are beyond capacity for the majority of institutions; and (3) there is little participation of stakeholders in decisions made concerning the use of assessment technology and opportunities to collect and use the learners' data. The researchers concluded that effective privacy protection in technology-enabled assessment requires a multidimensional approach balancing innovation with accountability, ensuring that surveillance infrastructure does not undermine educational equity while protecting vulnerable learner populations from data exploitation.

15:20:00 - 15:50:00

The Last Mile: What Happens After the Data Lands

Location: Trinity Ballroom - Salon 4

Speakers: Josh Linkletter

Educational assessment has spent fifty years getting the measurement right. The instruments are better than they have ever been. And yet the shared experience of ministries, examining bodies and researchers is that results land and very little changes. The report is published. The professional development day is held. And Tuesday looks like Monday. This session argues that the last mile is not a measurement problem. It is an ownership problem. Everyone in this field owns their own end of the work; almost no one owns the seam between the data being produced and the practice it was meant to inform. Offering a practitioner's perspective rather than a researcher's, the session focuses on one part of that seam: the difference between the measures organizations report and the measures they can act on. Nearly all formal assessment output is trailing — it describes learners who have already moved on. Closing the last mile means pairing it with a small set of leading measures that can be moved this week, each one owned by a named person and reviewed on a short cycle. The mechanics matter more than the concept. The idea of leading indicators has been in the education literature for two decades and still does not reach most schools, which suggests the gap is not one of knowledge but of operating discipline. The session also makes a case that will be unfamiliar to some: that a clear, well-chosen measure is not a constraint on professional judgement but a condition for it. Educators given a number and left to determine the method are more autonomous, not less. Attendees will leave with a practical method for identifying such measures in their own organization, whatever their role, and with a starting reference set built for education contexts rather than adapted from business.

16:00:00 - 16:30:00

Evaluating Technology-Enhanced Item Types for Science Assessments with Students with Significant Cognitive Disabilities

Location: Trinity Ballroom - Salon 3

Speakers: Dr. Melissa Gholson

Introduction: Accessible Teaching, Learning and Assessment Systems (ATLAS) evaluated five technology-enhanced item (TEI) types—drag-and-drop, drop-down (cloze), hot spot, table match (matrix), and simulation—for inclusion in science assessments with students with significant cognitive disabilities. This study investigated whether TEIs validly measure multidimensional science constructs for students with significant cognitive disabilities and whether TEI features introduce construct-irrelevant barriers that undermine trustworthy inferences. Methodology: Using a modified cognitive-lab protocol tailored for this population, researchers observed students across Complexity Bands 1–3 in two states as they completed five TEI types. Observational data were paired with teacher ratings of student performance and post-administration interviews with test administrators. Multiple data sources enabled a holistic appraisal of accessibility, presentation, engagement, and administration variability for each item type. Results: Three TEIs—drag-and-drop, hot spot, and drop-down —demonstrated the strongest potential for operational use. Drag-and-drop was effective when facilitating independent or minimally supported responses and aligning closely with common instructional activities. Hot spot and drop-down worked well but often required modeling. Drop-down items were text-heavy and navigation-dependent, increasing reliance on test administrators for reading and navigation support. Table

match (matrix) and simulation items presented substantial barriers. The matrix formats-imposed split-attention effects and working-memory demands, small touch targets reduced independence, and administration variability influenced outcomes. Simulations required multistep navigation, experienced technical inconsistencies (loading, rendering), and conflated interface navigation skill with science understanding. Conclusion: Findings indicate that TEI viability depends on alignment with students' instructional experiences and device interactions. Trustworthy and transparent assessment conclusions require TEI designs that minimize navigation load, provide clear visual and interactive cues, ensure consistent rendering across devices, and use appropriately sized touch targets and practice items aligned to instruction. Formats that impose unfamiliar interfaces or heavy scaffolding risk introducing construct-irrelevant variance and should be avoided or significantly redesigned. This study provides empirical evidence to guide defensible TEI selection, design, and administration in large-scale science assessments for students with significant cognitive disabilities, supporting valid, equitable, and transparent measurement. Attendees will gain actionable design and administration guidelines to enhance technology-enhanced items for assessment.

16:00:00 - 16:30:00

Universality, Not Uniformity: Lessons from Schools2030 on Implementing Digital Assessment Infrastructure in 10 Countries

Location: Trinity Ballroom - Salon 1

Speakers: Bronwen Magrath, Richard Bartle-Tubbs

This presentation examines the iterative development of a central digital assessment management system for Schools2030, a ten-year, ten-country global education programme. The objective of the assessment data management system is to transition from disparate, manual data silos—created by challenging and varied cultural contexts—to a unified, queryable digital environment that integrates student learning outcomes, classroom environment observations, and program implementation data. The system provides programmatic decision support by making complex assessment data accessible and actionable for country-level program teams. We employ a "Crawl-Walk-Run" implementation methodology, prioritizing sustainable, low-tech entry points before any attempts to scale to complex infrastructure. The approach utilizes a "Build-Measure-Learn" loop, starting with low-fidelity prototypes that build on 3 years of assessment tool development across 10 countries, representing diverse outcome measures, educational contexts, and languages. The prototype data management systems involve a mixture of manual and digital quality processes, which are field-tested with regional staff to gather feedback on usability and technical feasibility. Our discovery and implementation results highlighted that technical innovation must be secondary to human process, especially with varying levels of technology access, familiarity with assessment practices, and pedagogical norms across different countries. We found that standardizing core demographic elements, while ensuring flexibility for local assessment variations, is critical for maintaining oversight of a global model. A primary technical challenge identified was the manual effort required for data transformation; thus, we prioritized establishing a "Teacher ID" as the essential bridge to link disparate datasets. Furthermore, data quality assurance must be a collaborative process between global oversight and country-level validation rather than a purely automated function. Digital assessment environments offer the opportunity for a holistic profile of school performance, correlating teacher training attendance with student growth. However, the challenge lies in maintaining data integrity and accompanying quality processes across varied geographic contexts. The implication for educational assessment is that concrete value resides in simple, queryable visuals that solve immediate program needs. By focusing on interoperable interfaces and clear standard operating procedures, organizations can build sustainable data cultures that survive staff turnover and technical evolution.

16:00:00 - 16:30:00

Human–AI Collaborative Scoring in Resource-Constrained Contexts: Advancing Trust, Transparency, and Fairness in Educational Assessment in Sub-Saharan Africa

Location: Trinity Ballroom - Salon 4

Speakers: Prof. Mfonobong Umobong

Artificial intelligence (AI) adoption in educational assessment is receiving worldwide attention; still, it is not widely used in resource-limited settings, especially in Sub-Saharan Africa, as many issues of validity, fairness,

transparency, and stakeholder trust are still unresolved. In most of these situations, however, the process of assessment remains highly dependent on human judgment, particularly in constructed-response tasks, posingchallenges in scoring consistency, efficiency and scalability. The study examines the possibilities of human-AI collaborative scoring as a context-sensitive and scalable tool to improve the quality of assessments without replacing the standard human-centered practices. The study adopted a mixed methods design with data obtained from 500 undergraduate students undertaking courses in six universities from all geopolitical zones in Nigeria. Responses to a standardized constructed-response assessment are compared in three conditions: Human scoring, AI scoring; Hybrid model with human; and AI raters where the AI-generated scores are revised and corrected by trained human raters. Quantitative analyses examined inter-rater reliability, human versus AI agreement scores, consistency of scoring and efficiency across conditions. To complement the scoring analysis, survey data are gathered from key stakeholders, such as students, lecturers, and assessment practitioners, to examine the perceptions of trust, fairness, transparency, and acceptability of AI-assisted assessment. Such a multi-stakeholder approach allows understanding of the socio-technical aspects of AI integration into assessment systems in the region in a more holistic manner. The research is anchored in the concept of augmented validity, which adds to the conventional validity frameworks the transparency, auditability, and human control as the key elements of the AI-assisted assessment. Results are anticipated to demonstrate that although AI-only scoring is efficient, they could be constrained by interpretability and trust issues. Conversely, the hybrid model is expected to increase reliability of scoring and maintain human judgment, which will improve the technical quality and confidence of the stakeholders. By leveraging evidence based on various institutions in different regional settings, the study offers a sound empirical basis of the responsible adoption of AI into assessment with useful advice to balance innovation with trust in Sub-Saharan Africa. Keywords: Human–AI Collaborative Scoring, Automated Essay Scoring (AES), Augmented Validity, Fairness in Assessment, Sub-Saharan Africa.

16:00:00 - 16:30:00

Supporting Learning and Meeting Regulatory Assessment Requirements: Mathematics Teachers' Perspectives

Location: Trinity Ballroom - Salon 2

Speakers: Corrine Robinson, Dennis Alonzo

Preparation for high-stakes testing can constrain teachers' practice as they strive to meet conflicting expectations from stakeholders. Finding the balance between test preparation and deep learning is a challenge for teachers, particularly in the final years of schooling. We adopted a mixed-methods explanatory design within a socio-cultural constructivist paradigm to understand assessment practice in secondary mathematics, focusing on how teachers understand and navigate assessment requirements to support student learning. In stage 1, 72 Australian secondary mathematics teachers completed a self-report questionnaire. In stage 2, eight teachers participated in semi-structured interviews to provide deeper insight into the trends that emerged from stage 1. Results from the quantitative stage highlight the dominance of regulatory assessment functions, such as measurement, evaluation, and norm-referencing. The importance of regulatory assessment in informing students and parents about students' progress toward high-stakes testing was highlighted. Teachers described the scrutiny of parents, students and colleagues on assessment results and the impact this had on their classroom assessment practice. One such result of this scrutiny was frequent testing for reporting and ranking with fewer opportunities for responding to feedback, with teachers prioritising alignment with other classes over meeting needs identified from testing. Some teachers reported working in environments that were supportive of teacher professional judgement, and this resulted in the belief that they had greater autonomy to meet student needs in their classrooms. These teachers described innovative ways they enhanced their formal assessments to meet a wider range of purposes, including increasing student motivation, informing teacher planning, supporting student self-regulation, and improving learning. Our study highlights that when regulatory assessment is paired with strong professional support, teachers are more likely to view accountability as a shared responsibility rather than external surveillance. Trust between stakeholders is built when accountability systems are transparent, fair, and connected to improving student learning, rather than solely focused on compliance and reporting. Recognising teacher expertise and developing this through quality professional learning, strengthens assessment design capability and expands assessment practice repertoire to better support student learning. Ultimately, trust grows when stakeholders see that accountability mechanisms actively enable better teaching, richer feedback and improved student outcomes.

16:00:00 - 16:30:00

Effect of AI-Powered Reading Tools on the Literacy Development of Junior Secondary School Students in Abia State, Nigeria

Location: Carlton

Speakers: Dr. Kelechi Ihechu, Dr. Sara Hassan, Prof. Ursula Akanwa

Abstract: As artificial intelligence tools enter more classrooms, education systems increasingly need assurance that AI-generated feedback and scores are transparent, pedagogically sound, and fairly applied. This quasi-experimental study examined the effect of AI-powered reading tools on the literacy development of Junior Secondary II (JSS II) students in Abia State, Nigeria, and considers what the findings imply for trust and transparency in technology-enabled literacy assessment. Two research questions and two null hypotheses guided the study, together with a third, exploratory research question on implementation. From a population of 4,825 JSS II students across 269 public junior secondary schools (2024/2025 session), 125 students were purposively drawn from two schools, with one intact class from each school serving as the experimental (n = 62) or control (n = 63) group. The 30-item Literacy Skills Achievement Test (LSAT), comprising 15 reading-comprehension and 15 vocabulary items, was face- and content-validated by subject and measurement experts and yielded a Kuder-Richardson 20 reliability coefficient of 0.88 following a pilot test with 30 non-participating students in Owerri, Imo State. Ethical approval and informed parental consent and student assent were obtained prior to data collection. The experimental group used Google Read Along and Microsoft Reading Coach for 45 minutes, three times weekly, over 12 weeks; the control group received conventional print-based instruction on the same schedule. Data were analysed using means and standard deviations for the research questions and Analysis of Covariance (ANCOVA) for the hypotheses at ? = 0.05. Students who used the AI-powered reading tools recorded significantly higher post-test scores in reading comprehension, $F(1, 122) = 11.04$, $p = .004$, and vocabulary acquisition, $F(1, 122) = 11.36$, $p = .001$, than the control group. Based on the findings, it was recommended among others that teachers should receive structured training, prioritised ahead of wider adoption, on both the instructional use of AI reading tools and how to interpret the progress data these tools generated.

16:00:00 - 16:30:00

Evaluating Examination Access Arrangements for Candidates with Visual Impairments in the West African Senior School Certificate Examination in Ghana

Location: Bay

Speakers: Wendy Enyonam Addy-Lamptey

Examination access arrangements are provided by examination boards to ensure that candidates with special educational needs are not disadvantaged but are able to take examinations within their capabilities. The West African Examinations Council (WAEC) provides access arrangements for the candidates with visual, hearing and physical impairments, and other learning disabilities in the West African Senior School Certificate Examination (WASSCE) in Ghana. This study examines the access arrangements made for candidates with visual impairments in the form of large print and braille question papers, use of scribes (amanuensis), computers, screen readers, audio recorders and extended time. A descriptive survey was employed to evaluate the effectiveness of these arrangements, the resources required, and their ease of use. Data were collected from 57 participants, comprising current and past candidates with visual impairments, using questionnaires. In addition, semi-structured interviews were conducted with 10 qualified teachers for visual impairment from inclusive senior high schools and two (2) others. The quantitative data was summarised using frequencies and percentages and qualitative data analysed thematically. Findings indicated that large print question papers, use of computers and screen readers were rated as highly effective while challenges remained with braille provision and scribe use. One of the key recommendations from the study is that examination boards conduct regular needs assessments to ensure that access arrangements are appropriate, adequately resourced, and responsive to candidates' needs.

16:00:00 - 16:30:00
Student and Educator Voice in Assessment Design
Location: Trinity Ballroom - Salon 5
Speakers: Dr Bridget Mthembu
Title: Systems Thinking in Educational Reform: Stakeholder Engagement in South Africa's General Education Certificate (GEC) Implementation Educational reform is most effective when viewed as a dynamic system where multiple actors interact to shape outcomes. South Africa's General Education Certificate (GEC) implementation in South Africa exemplifies this systems theory approach, embedding bottom up stakeholder engagement within a broader framework of policy, curriculum, and assessment reform. The GEC, positioned at NQF Level 1, is designed not as an exit qualification but as a formative milestone at Grade 9, guiding learners into academic, vocational, or occupational pathways. Its success depended on the active participation of teachers, district officials, provincial education leaders, and communities, who collectively cocreated assessment practices and ensured contextual relevance. Working groups across provinces played a pivotal role in operationalizing the GEC assessment model, which integrates standardised curriculum tests, school based assessments, and inclination profiling. Teachers piloted integrated projects to capture 21st century skills, district officials coordinated moderation processes, and provincial teams adapted inclination assessments to diverse learner needs. This systemic interplay reinforced trust in the reform process, as stakeholders saw their voices reflected in both design and implementation. The systems theory perspective highlights how bottom up engagement was not isolated but interconnected with top down policy frameworks, creating feedback loops that strengthened legitimacy and transparency. Pilot findings revealed improved consistency across provinces, greater teacher confidence in alternative assessments, and enhanced parental trust through accessible reporting dashboards. By situating stakeholder engagement within a systems framework, the GEC pilot demonstrates how reforms can balance equity, credibility, and inclusivity. This paper argues that systems theory provides a valuable lens for understanding educational reform: stakeholder engagement functions as a critical subsystem that interacts with policy, curriculum, and assessment to produce sustainable change. The South African case offers lessons for other contexts, showing that embracing bottom up voices within systemic design builds trust, transparency, and resilience in assessment reform.

16:45:00 - 18:00:00
AGM (for IAEA Members)

18:30:00 - 20:30:00
Gala Dinner

Friday, October 02, 2026

07:00:00 - 08:30:00
Breakfast

08:00:00 - 10:00:00
Registration

08:30:00 - 09:15:00
Designing for Interpretation: How Interactive Assessment Reporting and Professional Learning Support School Improvement Processes
Location: Trinity Ballroom - Salon 3
Speakers: Amie Presley, David Cameron, Janet O'Reilly, Stefanie De Jesus
The Research Department serves as a central source of evidence and analysis for the Toronto District School Board (TDSB), supporting Board and school-level improvement planning. Based on twenty-five years of research, the department has identified a set of evidence?based indicators associated with secondary school graduation and postsecondary success. These indicators form the foundation of the department's analysis and reporting work. This presentation examines how the reporting and use of assessment results for school improvement planning (SIP) can be strengthened through the integration of interactive visual reports of curated evidence-based indicators and sustained professional learning focused on data literacy and responsible data use. Drawing on their work at the TDSB, members of the Research team will describe their approach to building data literacy and supporting principals' interpretation of assessment, achievement and contextual data. The iterative nature of the work will be highlighted showing how feedback from practitioners, alignment with system priorities and provincial reporting requirements, and advances in technology continue to inform the refinement of both the reporting/analytical tools and the professional learning strategies. In September 2023, the department began transitioning from static tabular reports to interactive data reporting using Power-BI. The shift began with perception data from student surveys, but over time, the evidence-based indicators were integrated to support more meaningful interpretation of assessment, achievement, and contextual data. During the 2024–2025 year, efforts focused on improving access to the new reports and working with Superintendents to support system?level planning. In the following year, the work expanded when the department was invited to engage with school principals, which led to pairing the interactive reports with professional learning to support principals' data literacy and application of assessment evidence in SIP. Feedback indicated that while professional learning improved clarity and access, leaders required repeated opportunities to engage with data, ask questions, and connect findings to SIP. In response, a three?part professional learning series is planned for the upcoming school year, along with an emerging evaluation framework that will evaluate the implementation and effectiveness of this model. The session illustrates how clarity, accessibility and responsible communication in assessment reporting are not only features of report design, but key outcomes of a deliberate alignment between data representation, professional learning, and system context. Assessment systems benefit from grounding indicators in evidence linked to meaningful long term outcomes and from aligning interactive reporting with sustained professional learning that supports data literacy and interpretation.

08:30:00 - 09:15:00
Mapping Accessibility and Inclusion Practices in Large-Scale Student Assessments: International Patterns, Shared Challenges, and Future Directions
Location: Trinity Ballroom - Salon 4
Speakers: Elodie Vezon
Ensuring that large-scale student assessments are accessible and inclusive has become an increasingly important challenge for assessment organizations worldwide. As student populations become more diverse and expectations around inclusion continue to grow, institutions involved in large-scale assessment are under increasing pressure to address accessibility, accommodations, and participation in ways that are both educationally meaningful and methodologically robust. Yet international visibility on how these practices are currently defined, organized, and implemented remains limited. This presentation reports on the results of an

international mapping survey that I developed and am leading through the FLIP+ association. The survey was designed to document current practices related to accessibility, inclusion, accommodations, exclusion policies, internal expertise, organizational responsibilities, and institutional needs in large-scale student assessments. The presentation has three main aims: First, it provides a comparative overview of how assessment organizations and related institutions currently approach accessibility and inclusion. Second, it identifies recurring challenges and areas of convergence across contexts, including the distribution of responsibilities, the availability of expertise and guidance, and the difficulty of translating inclusive intentions into operational assessment practices. Third, it examines how these findings can support the development of shared directions, criteria, and practical resources for strengthening inclusive assessment work internationally. Beyond describing current practices, the survey findings point to a broader need for more structured reflection on how inclusive assessment practices are defined, supported, and evaluated across contexts. They also highlight that, despite important differences in national and institutional settings, organizations face several common questions and unmet needs. In this respect, the presentation argues that international dialogue is valuable not only for documenting diversity, but also for clarifying shared priorities and supporting collective capacity-building. This proposal contributes to the conference subtheme of Transparency by making current practices more visible, clarifying shared institutional challenges, and supporting more explicit, coherent, and informed approaches to inclusive large-scale assessment.

08:30:00 - 09:15:00

Learning Maps in Assessment Systems: Balancing Accountability and Instructional Support

Location: Trinity Ballroom - Salon 5

Speakers: Brooke Nash, Russell Swinburne Romine, Vanessa Simmering

In the U.S., states are exploring how large-scale, traditional end-of-year assessments can be reimagined. Flexibility from the U.S. Department of Education allows for implementation of through-course models that yield summative results for accountability purposes, but also information that supports instructional decision-making throughout the year. This has the potential to reduce testing fatigue, strengthen instructional alignment, and produce data that can be used for both classroom and policy needs. To ensure quality of assessment design and aligned instructional decisions, we argue for the use of learning maps as an architecture for developing instructionally embedded assessment systems. Learning maps are large, fine-grained, intentionally inclusive, research-based cognitive models that reflect the diversity of how learners may acquire knowledge, skills, and understandings in a domain (Swinburne Romine et al., 2025). By specifying multiple entry points, varied learning trajectories, and alternative pathways through content with potential barriers to accessibility, learning maps are designed to support all students regardless of differences in prior knowledge, rates of learning, and accessibility needs. As a result, learning maps provide a foundation for informative assessments that allow diverse learners to demonstrate growth over time. Large-scale assessments should be grounded in theories of action that clarify the purposes and uses of their results (Marion et al., 2016). Using learning maps for an assessment architecture ensures transparency and coherence across the system by formalizing claims about the relationships among assessment, learning, and instruction within the theory of action. We present two case examples of learning map-based instructionally embedded assessments that support accountability reporting and instructional decision-making:The Dynamic Learning Maps (DLM) alternate assessment system assesses students with significant cognitive disabilities in English language arts, mathematics, and science in Grades 3 through 12. Pathways for Instructionally Embedded Assessment (PIE) is an innovative assessment model that allows Grade 5 students to demonstrate competencies in mathematics as they learn while also producing results for summative assessment purposes. These examples illustrate how learning maps can support a coherent assessment architecture for both valid accountability reporting and instructionally meaningful uses of assessment data, strengthening trust in the results and alignment across the system.

08:30:00 - 09:15:00

Determinants of Performance on the SweSAT

Location: Trinity Ballroom - Salon 2

Speakers: Erika Majoros

Large and persistent score differences between men and women have been observed in the quantitative

section of the Swedish Scholastic Aptitude Test (SweSAT), differences that cannot be fully explained bydifferential item functioning (DIF) (Wedman, 2017). Previous fairness research on the SweSAT has therefore largely focused on gender-related DIF, consistently concluding that only a modest share of the observed score gap can be attributed to item bias and that the test functions similarly for men and women overall (Bränberg et al., 1990; Wedman, 2018). In contrast, comparatively little research has examined how broader contextual and background characteristics shape performance on the SweSAT, despite its central role in higher education selection and its implications for equity and access. This study addresses this gap by investigating determinants of performance in the quantitative section of the SweSAT using data from the last pre-pandemic spring administration (n = 48,788), linked to rich register-based background information from the Gothenburg Educational Longitudinal Database. Framed within an equity and fairness in assessment perspective, the study examines how social and educational contexts contribute to performance differences across groups of test takers. The analysis proceeds in two steps: Confirmatory factor analysis is used to examine the internal structure of the quantitative section and identify substantively meaningful subscales. Mixed-effects item response theory models are applied to estimate group differences in latent ability across these subscales while accounting for item-level variation. Gender, age, parental education, and language background are included as predictors. These findings contribute to a broader understanding of fairness in large-scale admissions testing by demonstrating how observed performance differences may reflect unequal learning opportunities and accumulated educational advantages and disadvantages. From an assessment policy perspective, the results underscore the importance of complementing traditional DIF-based fairness evaluations with analyses that explicitly consider the broader social context in which high-stakes tests operate. References: Bränberg, K., Henriksson, W., Nyquist, H., & Wedman, I. (1990). The influence of sex, education and age on test scores on the Swedish Scholastic Aptitude Test. Scandinavian Journal of Educational Research, 34(3), 189–203. https://doi.org/10.1080/0031383900340302 Wedman, J. (2018). Reasons for gender-related differential item functioning in a college admissions test. Scandinavian Journal of Educational Research, 62(6), 959–970. https://doi.org/10.1080/00313831.2017.1402365

08:30:00 - 09:15:00

Towards Transparent Admissions Decisions: Comparing IELTS, TOEFL, and Duolingo as Predictors of Academic Success

Location: Trinity Ballroom - Salon 1

Speakers: Mehrdad Yousefpoori-Naeim, Samira ElAtia

This study investigates transparency in educational assessment by analyzing the use of standardized English proficiency tests in admissions decisions and the communication of their predictive value to stakeholders. Institutional data from 1,569 international students at a large Canadian university are used to compare the predictive validity of IELTS, TOEFL, and the Duolingo English Test (DET) with respect to cumulative GPA over a 4 year period. The study uses both traditional statistical analyses and machine learning methods for promoting transparent reporting of assessment outcomes. Pearson correlations, regression models, and interpretable tree-based machine learning algorithms were used to examine the relationship between overall test scores, subscores, and academic performance. Across all three tests, overall scores explained only a small variance in GPA (R^2 ? .03-.04), which highlights the need for cautious and responsible use of test scores in high-stakes admissions decisions. A key finding of the study is that reading subscores emerged as the most robust and consistent predictor of academic success for all tests. IELTS reading particularly showed the strongest relationship with GPA. For TOEFL, reading was the strongest predictor as well, with a smaller contribution from speaking. The integrated subscores in DET showed limited explanatory power compared to the other two tests. These findings enhance transparency in assessment design and scoring by identifying which components of language proficiency meaningfully relate to academic outcomes. The study has three main implications for educational assessment:The study supports clearer evidence-based communication of the strengths and limitations of English proficiency scores in admissions. It has more transparent and fair decision thresholds based on subscores (rather than overall scores). It raises important considerations for equity when alternative language tests are treated as equivalent in university admissions. The study contributes to ongoing discussions about openness, fairness, and accountability in assessment practices within higher education.

09:15:00 - 10:00:00

From Legacy to Digital at Scale: Building Trust and Transparency in Alberta's Assessment Transformation

Location: Trinity Ballroom - Salon 5

Speakers: Charles Anifowose, Pascal Couture

Over the past decade, Alberta Education has undertaken an ambitious large-scale assessment modernization initiative. Serving approximately 700,000 students annually, the province transitioned from a predominantly paper-based system to a fully integrated digital assessment ecosystem capable of supporting tens of thousands of concurrent users across diverse devices and contexts. This presentation examines the practical realities of implementing digital assessment at scale, focusing on how technology can strengthen trust and transparency while addressing real-world constraints. Beginning with a complex legacy environment characterized by fragmented systems, paper-based workflows, and limited concurrency, Alberta's journey required not only technological innovation but also careful attention to stakeholder confidence, operational continuity, and equitable access. We will highlight key design principles that guided the transformation, including cloud-based architecture, device-agnostic delivery, human-centred design, and security by design. Particular attention will be given to how the solution supports high-stakes assessments with robust reliability, scalable concurrency, and consistent user experience across PCs, Macs, Chromebooks, and tablets. The session will also explore challenges encountered along the way, including managing the transition from paper to digital, ensuring accessibility for all learners, maintaining validity and reliability in new formats, and supporting educators and administrators through change. Lessons learned will be shared on governance, phased implementation, stakeholder engagement, and aligning technology decisions with policy and assessment objectives. Now fully implemented, Alberta's digital assessment ecosystem has improved efficiency, reduced turnaround times, expanded access to online testing, and enhanced data availability for educators and policymakers. More importantly, it demonstrates how large-scale digital transformation, when executed thoughtfully, can reinforce trust in assessment systems while increasing transparency and resilience. This case study offers practical insights for jurisdictions seeking to modernize their assessment systems, balancing innovation with responsibility in an increasingly digital world.

09:15:00 - 10:00:00

Making Learning Visible: Assessment Practices that Support Data?Informed Instruction

Location: Trinity Ballroom - Salon 4

Speakers: Betty Farrell, Lorie Mills, Marylin Dawson

Trustworthy assessment systems are grounded in transparency, curriculum fidelity, meaningful evidence of learning, and shared understanding between educators and students. At Wellington Catholic District School Board (WCDSB), intentional system leadership serves as the foundation for establishing these conditions. This session reflects an early stage of a system-wide journey focused on clarifying instructional priorities in literacy and numeracy, aligning assessment supports, and building coherent expectations across schools to promote consistent, high-quality classroom practice. The session highlights WCDSB's leadership approach to designing coherent structures that support effective teaching, learning, and assessment in literacy and mathematics. System leaders provide educators with aligned tools and professional learning to support the purposeful use of assessment evidence, including early screeners, pre-assessments, and progress checks. This evidence is used to develop a shared understanding of student strengths, needs, and priority learning gaps, guiding instructional decisions rather than functioning as isolated measures. Curriculum-aligned scope and sequences for mathematics anchor instruction and assessment across the system. These resources articulate clear learning progressions, instructional priorities, and assessment checkpoints that support consistent implementation. Board-approved digital tools, such as Edwin, are embedded at the point of need to ensure equitable access to instructional and assessment resources aligned with provincial curriculum expectations. Within this framework, educators use professional judgment to plan responsive instruction informed by assessment evidence. Assessment functions as a tool for learning, supported by system-designed structures that make student learning visible, transparent, and actionable. Participants will explore how these assessment practices build trust and transparency around learning expectations and outcomes, helping students understand what they are learning, why it matters, and how they can improve. The session also demonstrates how educators use assessment evidence to inform instruction, respond strategically to learning needs, and

uphold equity and fairness in classroom assessment. Drawing on system-level literacy and numeracy practices, presenters will illustrate how clear data views support professional dialogue, strengthen instructional decision making, and balance accountability with instructional support. This session aligns with IAEA priorities related to responsible data use, student and educator voice, and trust in assessment practices.

09:15:00 - 10:00:00

Between Automation and Responsibility: The Conditions for Trust in AI-Based Assessment

Location: Trinity Ballroom - Salon 2

Speakers: Jean-Philippe Rivière

What if, for a moment, we trusted artificial intelligence to take on part of the tasks required for designing and administering digital assessments? Implementing a digital assessment relies on a complex chain of operations involving didactic, pedagogical, editorial, docimological, statistical, and technical dimensions, alongside organizational constraints and issues of social acceptability. While the prospect of fully AI-driven assessment based on artificial general intelligence remains out of reach, recent advances already enable the integration of AI systems into existing frameworks. In this context, an “agent-based” approach offers a pragmatic pathway: specialized agents, relying on learning models, can operate at different stages of the process to improve efficiency and quality, while remaining embedded within a deterministic infrastructure that ensures system robustness. These agents can support test construction by selecting or generating items from metadata-enriched item banks, as well as by controlling their quality (consistency, level of difficulty, contextual diversity). They can also anticipate the equivalence between different versions of a test prior to psychometric validation. One of the most significant contributions of AI concerns the scoring of constructed responses. While closed formats such as multiple-choice questions have well-known limitations, current systems are now able to reliably assess short responses and are beginning to handle more complex productions. This development paves the way for more authentic assessments, focused on reasoning and production, while reducing the constraints associated with continuous item renewal. However, significant challenges remain, particularly regarding data security, model control, and technological sovereignty, which argue in favor of local and controlled solutions. In the longer term, the orchestration of multiple agents could enable the supervision of the entire assessment process. In this perspective, AI appears less as a substitute than as a means of augmenting human capabilities, inviting a re-examination of the balance between automation and responsibility.

09:15:00 - 10:00:00

Building Confidence in Online Assessment Through Deliberate, Phased Design

Location: Trinity Ballroom - Salon 1

Speakers: Dr. Charles Mayenga, Gordon Chan

Theme Alignment: Digital Assessments: Opportunities and Challenges: Examining innovation, accessibility and practical implementation in digital environments. The Canadian Securities Institute (CSI) undertook a deliberate, phased transition from proctored paper?and?pencil examinations to fully online assessment delivery. To minimize risk and preserve score meaning, CSI first migrated hundreds of examinations to fixed, published online forms before progressing to dynamic item pools. This staged approach aligned with best practices articulated in the Standards for Educational and Psychological Testing (AERA, APA, NCME), emphasizing validity, fairness, and score comparability during periods of change. The initial publication of fixed forms served as a critical bridge between paper and digital delivery. By holding test content, structure, and form-level specifications constant, CSI was able to isolate and evaluate delivery?mode effects while preserving established pass-fail standards. Item development, form assembly, and review processes were standardized and documented to ensure consistent content coverage, cognitive complexity, and alignment to learning outcomes. Statistical monitoring at the form level supported ongoing evaluation of reliability, item functioning, cohort comparability, and test experience surveys providing empirical evidence that transitioning to online delivery did not introduce construct?irrelevant variance. Once equivalence and operational stability were established, CSI expanded to dynamic item pools designs to improve security, scalability, and scheduling flexibility. Through data analysis and governance procedures that prioritized an evolving perspective on test security risks, pools were monitored for exposure and content leaks, and new items that were added ensured that continuous testing remained appropriately covered to content, comparable and defensible. These controls reflected best practices that flexibility in administration must be counterbalanced by systematic design and

monitoring to preserve validity. Validity evidence was treated as cumulative and ongoing. CSI embedded numerous pre- and post- publishing quality checks, routine psychometric review cycles, integrating statistical analyses with stakeholder feedback to inform continuous improvement. By sequencing fixed forms before dynamic pools, and grounding each phase in recognized testing standards, CSI demonstrated a defensible pathway for modernizing professional banking and finance assessments while maintaining confidence in score interpretation and decision making.

09:15:00 - 10:00:00

Can an AI Examiner Be Fair? Designing, Validating, and Trusting Automated Oral Assessment at Scale

Location: Trinity Ballroom - Salon 3

Speakers: Sanketa Anand

The oral examination has been a cornerstone of educational and professional assessment for centuries — valued precisely because it cannot be faked in the way a written response can. It demands real-time reasoning, verbal articulation, and the ability to respond to probing follow-up. It is also expensive, difficult to standardize, and impossible to deliver at the scale that modern credentialing and educational assessment programs require. AI-driven conversational assessment is beginning to change that equation. Advances in large language models, automatic speech recognition, and real-time natural language understanding have made it technically feasible to deploy AI examiners capable of conducting structured oral assessments, evaluating spoken responses for content accuracy and reasoning quality, and adapting follow-up questions based on candidate answers — at scale, consistently, and without the scheduling and cost constraints of human examiners. But technical feasibility and assessment validity are different things. This presentation examines three foundational questions that any program considering AI oral assessment must honestly confront: Can an AI examiner assess fairly? Spoken language is profoundly shaped by accent, dialect, speech rate, and sociolinguistic background. Automatic speech recognition systems have well-documented performance disparities across demographic groups, and these disparities translate directly into scoring inequity if left unaddressed. How do you validate AI oral assessment for high-stakes use? The evidence standards required — construct validity, inter-rater reliability, standard-setting, bias analysis — are the same as for any assessment, but the methodology for meeting them in an AI context is still being established. How do you maintain candidate and institutional trust in a system where the examiner is a machine? Drawing on deployment experience from AI-assisted oral and conversational assessment programs across language proficiency, professional certification, and higher education contexts, this presentation provides an evidence-grounded answer to each question.

09:15:00 - 10:00:00

From Data Production to Policy Use: The Case of PASEC in Francophone Sub-Saharan Africa

Location: Foyer

Speakers: Aissata Foune Cisse

This literature review is situated within the broader context of evidence-based education policies, particularly the use of large-scale learning assessment data in policy decision-making. In Francophone Sub-Saharan Africa, the Programme d’analyse des systèmes éducatifs de la CONFEMEN (PASEC) constitutes a central mechanism for producing comparable data on student learning outcomes. However, the production of such data does not guarantee their effective use in policy decision-making processes. PASEC data take multiple forms, including published reports, raw datasets, and numerous secondary studies produced by researchers and institutions based on these data. This review identifies documented cases of policy use of these different forms of data and provides a critical analysis of their dissemination and appropriation dynamics. It highlights key gaps in the literature, particularly the limited knowledge regarding the processes, actors, and conditions that shape data use in policy decision-making, thereby contributing to advancing understanding of this emerging field of research. The frameworks of evidence-based policy and sense-making are mobilized to analyze these dynamics and to identify relevant avenues for future research and reflection, particularly in relation to the dissemination and appropriation of results. Special attention is given to the balance of roles and responsibilities between CONFEMEN and participating countries.

09:15:00 - 10:00:00

Learning from Mistakes: Using Digital Numeracy Screeners to Identify Student Misunderstandings of Core Mathematical Concepts

Location: Foyer

Speakers: Michael Slipenkyj

Introduction: Digital numeracy screeners have several obvious benefits, including control of testing time and automatic scoring. We show that digital screeners can also give meaningful insight into students' misunderstanding of fundamental mathematical concepts. Methodology: A large sample of Grade 4 students (n1000) completed the Digital Early Math Assessment@School (DEMA), a teacher-administered numeracy screener. We focussed on three numeracy domains: number writing, arithmetic equations, and fraction knowledge. Results: Three key error patterns were identified:Place-value transcoding errors were evident, that is, writing the number "1100060" for the auditory prompt "eleven-thousand and sixty".Students made errors showing an incomplete understanding of mathematical inversion (e.g., failing to recognize "+2" and "-2" cancel out when answering 8+2-2). Students incorrectly treated fractions as natural numbers (e.g., answering that 2/9 is greater than 1/3 because 2 and 9 are larger). Discussion and Conclusions Our results show that digital numeracy screeners can demonstrate students' misunderstanding of core mathematical concepts in real-time and at large-scale. Identifying which students make these errors and when different errors occur may help inform classroom practices and subsequently improve student outcomes. The challenge of how best to provide this information to teachers is an ongoing area of research.

09:15:00 - 10:00:00 | AI-Assisted Item Development in Practice: Lessons Learned from a Human-Led Pilot with External Subject Matter Experts

Location: Foyer

Speakers: Michelle McCoy

There has been increased interest in the use of generative artificial intelligence (AI) in educational assessment to improve item development processes. As test developers and vendors adjust procedures to include AI-assisted item writing, implementation within high-stakes environments must maintain focus on alignment, validity, and fairness. This poster session presents findings from a human-led pilot in which AI was incorporated as one tool within an established item development workflow involving external subject matter experts. Writers used publicly available AI tools in an iterative drafting process supported by structured workflows and multi-stage human review. All items were reviewed for content, alignment, and fairness. Outside reviewers were not informed of AI use. Efficiency gains were not realized in this implementation, despite being a primary goal. The process required substantial iteration, detailed prompting, and increased oversight. Variability in item writer preparedness, including writing and use of AI revealed the need for clearer expectations and targeted training. The key insight from this pilot is that AI functions most effectively as a tool within a human-led process rather than as a replacement for expert input or judgment. This work highlights the importance of structured implementation and clear boundaries for AI-assisted item development in high-stakes assessment contexts.

09:15:00 - 10:00:00

Research-Based Approach for Evaluating and Building Numeracy Skills in Digital Environments

Location: Foyer

Speakers: Ajay Karat, Dr. Robin Sharma

Early numeracy development is foundational to children's long-term mathematical success. Yet, assessment and instructional approaches in digital environments often lack the depth necessary to meaningfully support young learners. This work describes the design and development of a research-based early numeracy framework intended to assess and build mathematical understanding in young children through digitally delivered tasks. The framework draws on foundational research by describing early numeracy beyond numbers and arithmetic. It also includes algebraic thinking, data and probability, and spatial reasoning. The framework is designed as a multi-dimensional model in which each level reflects increasing cognitive complexity across age groups. Interactive tasks align with Canadian curriculum expectations while also reflecting how children naturally think, respond, and learn within digital environments. Grounded in math cognition theories (e.g., embodied cognition), these tasks serve a dual purpose: assessing children's current understanding and

developing more advanced conceptual knowledge. The resulting node-based system is adaptable andscalable. It can deliver progressively complex, research-informed assessment tasks based on each child's level. It can be aligned with provincial curricula and standards across educational contexts. Developing an early numeracy assessment framework is a complex process that requires many careful and informed decisions, as illustrated in this work.

09:15:00 - 10:00:00
Classroom Assessment for Student Well-Being: The Role of Trust, Respect, and Fairness
Location: Foyer
Speakers: Sumaiya Khanam Chowdhury
This study explores the underexamined conceptual connection between classroom assessment (CA) and student well-being (SW) from a global expert perspective. Using a Classical Delphi (CD) approach, 51 experts from 15 countries participated in the first phase of the study, the qualitative phase. Findings revealed that experts agreed that CA affects SW, although the specific nature of this relationship requires further empirical investigation. Among the positive practices identified, trust-building, respect, and fairness were seen as highly influential. Assessment practices that provide student voice, choice, privacy, and opportunities for revision communicate that assessment is intended to support learning rather than merely judge performance, thereby fostering trust, mutual respect, and fairness. In this way, students are more likely to feel valued, emotionally safe, and respected by both teachers and peers. Conversely, assessment practices that are punitive, overly competitive, unclear, or narrowly focused on grades can undermine trust, weaken classroom relationships, and reduce students' sense of belonging and confidence.

10:00:00 - 10:25:00
Coffee Break with Exhibitors
Location: Foyer & York AB

10:25:00 - 11:10:00
Trusting times? Understanding assessment and the aims of education
Location: Grand Ballroom
Speakers: Professor Mary Richardson
We are living and learning in a time of both precarity and great possibilities, therefore it is incumbent upon those of us who work, plan, and research in educational settings to do all we can to promote and sustain an environment that learners can trust and within which they feel they belong. Globally we are emerging from a pandemic and are, as yet, unclear about its long-term impact on education. More recently new technologies, specifically GenAI have launched into educational settings and seem set to change how we think about many aspects of the learning experience from content generation to the validity of assessment practice. All of these challenges and opportunities reflect the patterns of change and continuity we have come to expect in education, and they emphasise the 'messy' nature of what it means to be a learner and to learn. In this presentation, I will draw on my work focused on the nature of trust and confidence in assessment and consider how this now sits alongside a public perception of new technologies and how these are viewed with scepticism and concern. We have the skills, competence and knowledge to challenge fearful views and I will argue that our priorities should include good public communication about assessment. Learners, of all ages, deserve to feel that their endeavours are valued and valuable because as Peters (1967) claimed [education] is not about the destination, but about "… travelling with a better view".

11:10:00 - 12:10:00 | **Closing Ceremony**
Location: Grand Ballroom

12:10:00 - 13:10:00 | **Lunch**
Location: Grand Ballroom