e-assessment community

8TH FLIP+ ANNUAL,EVENT ARNHEM, 18TH – 20TH JUNE 2025

Venue: Hotel Haarhuis, Stationsplein 1, 6811LG, Arnhem

AGENDA

Day 1 will start with welcomes, then we will hear from member institutions and associated contributors who will share their e-assessment experiences. In the afternoon, our host, the CitoLab, will explore the synergy between cutting-edge technology and educational assessment. This will be followed by an interactive presentation of the developments to date of the FLIP+ Library. This will include a demo of the library and presentations by the FLIP+ working Groups.

Day 2 will include three interactive themed sessions, followed by a closing session to wrap the activities of the whole event, which will end at 16:00.

See pages 3–9 for details on the scheduled sessions and pages 10–13 for details of presenters.

Wed 18th June

14:00 – 16:30 *FLIP+ Working Groups meeting (reserved for FLIP+ Working Group members only)*
18:00 – 18:30 On-site registration and collection badges
18:30 – 20:00 Welcome reception supported by Cito, at Cito, Amsterdamseweg 13, 6814CM, Arnhem

DAY 1: Thu 19th June

08:30 onwards On-site registration and collection badges

09:00 – 09:30 Welcome address

⇒ To Cito – *Saskia Wools, CEO, Cito*
⇒ To FLIP+ Annual event – *Thierry Rocher, President of FLIP+*
⇒ To the event programme – *Amina Afif, Executive Secretary, FLIP+*

09:30 – 10:30 Member presentations

NE: *Dutch Board for Tests and Examinations (CvTE):* ***Transitioning towards a renewed digital examination ecosystem*** *Hugo Hopstaken*

ES: *The National Institute for Educational Assessment* - ***Piloting the assessment of linguistic communication, STEM, digital, and multilingual competencies,*** *Alberto Díaz-Roncero Canales, Sabrina Gallego Verdi*

NO: *NDET* - ***From Pen and Paper to Keyboard and Spell Checker: An Innovative Journey for Foreign Language Exams in Norway,*** *Kim Buxton*

10:30 – 11:00 Group Photo and Coffee break

11:00 – 12:30 Member presentations (*contd.*)

UK: *AQA - Alphaplus* - ***Numeracy for real life: trialling an innovative new digital numeracy assessment and teaching and learning app with students and teachers,*** *Laura Verdasco Menendez, AQA, Hannah Rowe, Alphaplus*

MO: *INE-CSEFRS* – ***PNEA 2025: National Digital Assessments at Scale,*** *Noura Kelloul*

BR: *CAED* – ***Building Assessment and Monitoring Platforms for different Brazilian states: from Test administration to Feedback,*** *Carlos Palacios*

BE: *Agency for Higher Education, Adult education, Qualifications & Scholarships* (AHOVOKS) – ***Innovating Language Assessment: Digital testing at the Flemish Examination Board,*** *Heidi Dreesen and An Schotte*

12:30 – 13:30 Lunch break

13:30 – 15:00 **Innovation at CitoLab**

NE: *CitoLab Presenters - Dr. Joost Kruis, Ingrid Williams, Patrick de Klein, Joey Stuiver, Nafsika Lachana*

Supporting Educational Measurement with Edtech and AI

15:00 – 15:30 Coffee break

15:30 – 17:00 **FLIP + Working Groups and FLIP+ Library Update**

Facilitators: **FLIP+ teams** - *Amina Afif, Thierry Rocher, Charlotte Gill-Sotty, François-Xavier Cannes and members of the FLIP+ Working Groups*

Launch of FLIP+ Library project and a demo of the FLIP+ Library platform

17:00 **End of DAY 1 sessions**

17:15 – 22:00 Tour of the Open Air Museum in Arnhem and dinner in one of the old buildings

- **Nederlands Openluchtmuseum**

DAY 2: Fri 20th June

09:00 – 10:30 **Digital Assessments and Implementation**

*CUPA **(UK)** – Cambridge University Press and Assessment (UK) Presenter: Sanjay Mistry*

Are Schools Ready for Digital Exams? – Using a framework to evaluate international digital readiness?

*NDET **(NO)** Presenter: Kevin Steinman*

Digital high-stakes English exams in Norway – what have we gained?

*Vretta **(CA)** Presenter: Vali Huseyn*

Piloting a quality audit framework for technology-based assessment production

10:30 – 11:00 Coffee break

11:00 – 12:30 **Digital assessment and Process Data**

*CUPA **(UK)** – Cambridge University Press and Assessment (UK) Presenter: Abdullah Ali Khan*

Supporting Students' Reasoning – Developing A Digital Annotations Toolkit for High-stakes Exams

*DEPP **(FR)** Presenter: Aurelie Lacroix, Franck Salles*

A case study using Process Data to enhance the Interpretation of Assessment Outcomes

*AQA Global Assessment Services **(UK)** Presenters: Shaun Crowley and Simon Trevers*

A practical view of modularising the end-to-end process of e-assessment development and delivery

12:30 – 13:30 Lunch break

13:30 – 15:00 **AI for Better Digital Assessments**

*AQA **(UK)** Presenters: Jiazheng Li, PhD student Kings College London, Cesare Aloisi, AQA*

New Age for Digital Assessment: Explainable Scoring with Large Language Models

*CAEd **(BR)** Presenters: César Soares, João Freire, Luiz Ribeiro*

AI-Driven NLP Analysis of Constructed Response Items in Teacher Questionnaires: A Pilot Study

*CITO **(NE)** Presenters: Tijmen Poell & Marica Balk*

SchrijfBlik: Safeguarding the validity of writing assessment in the age of AI

15:00 – 16:00 **Closing session and end of event**

ABSTRACTS

DAY 1: MEMBER PRESENTATIONS AND THE FLIP+ LIBRARY

Dutch Board for Tests and Examinations - CvTE (NE)

Title: Transitioning towards a renewed digital examination ecosystem

Presenter: Hugo Hopstaken (College voor Toetsen en Examens)

While Dutch education is rapidly digitalizing, central exams remain largely paper based. This discrepancy not only affects student motivation, it also limits the growing demand to digital assess applied skills and accommodate special needs effectively.

The Dutch Board for Tests and Examinations will present **the vision and strategy for transitioning towards a renewed digital examination ecosystem over the next five years**. This transformation will be a collaborative effort **involving schools, teachers, and key institutions** such as CvTE, Cito, DUO, and the Ministry of Education. The program's goals include introducing digital modules for examination of applied mathematics, listening skills and tools for assessing written work for language subjects, improving the exam process and logistics, enhancing the efficiency of grading exams and introducing more flexibility in exam scheduling. We will explore **emerging technologies like AI** for exam construction and correction and a sustainable digital infrastructure will be developed.

The National Institute for Educational Assessment - INEE (ES)

Title: Piloting the assessment of linguistic communication, STEM, digital, and multilingual competencies

Presenter: Alberto Díaz-Roncero Canales, Sabrina Gallego Verdi

The National Institute for Educational Assessment (INEE) is the body of the Ministry of Education, Vocational Training and Sports responsible for evaluating the Spanish education system. One of its missions is to develop multi-year plans for the general assessment of the education system and to establish methodological and scientific standards that ensure the quality, validity and reliability of educational assessments.

The current Spanish education law provides for a **general evaluation of the education system**, conducted on a sample-based and multi-year basis, in the sixth year of primary education (sixth grade) and the fourth year of compulsory secondary education (tenth grade).

The first edition of the 6th-grade Primary Education assessment is scheduled for the 2025-2026 school year. The following competencies are expected to be evaluated: linguistic communication, STEM, digital, and multilingual. The assessment will be administered entirely in digital format and in all five co-official languages. Between May and June 2025, the fieldwork for the pilot assessment of multilingual and digital competencies will be conducted.

Our presentation aims to provide **a brief theoretical framework on how these competencies were assessed in the pilot test and to showcase examples of digitized assessment units for multilingual and digital competencies.**

AQA - Alphaplus (UK)

Title: Numeracy for real life: trialling an innovative new digital numeracy assessment and teaching and learning app with students and teachers

Presenter: Laura Verdasco Menendez, AQA, Hannah Rowe, Alphaplus

AQA have been developing **an innovative new digital numeracy assessment** with an associated **teaching and learning (T&L) app** to help improve the financial numeracy of students in England. The proof-of-concept phase of this product development involved a large-scale trialling project, with over 1,400 prototype assessments sat by students in the autumn of 2024. Feedback was sought from teachers and students on the vital elements of the product and T&L app to inform future development.

This paper will examine **the motivations for the trial and the approach taken**, as well as data on item difficulty and which topics performed well/poorly to inform development. **Data from both the qualitative and quantitative strands**

to the trial will be presented. We will also cover how focus groups (N = 8 groups, 38 learners), teacher surveys (N = 18) and student surveys (N = 611) were used to get feedback on the assessment and the T&L app.
The findings from the **trial data** suggest that the app was seen as useful, and there was enthusiasm for a qualification assessing numeracy in real-life contexts. However, further work is needed to refine item demand and context within this new approach to both assessment and T&L in England.

INE-CSEFRS (MO)
Title: PNEA 2025: National Digital Assessments at Scale
Presenter: Noura Kelloul

The National Student Achievements Assessment Program (PNEA) in Morocco reached a key milestone in 2025 with **the implementation of a fully online, computer-based national assessment** using the TAO platform. Following earlier cycles in 2008 (paper-based), 2016, and 2019 (digitally supported), the 2025 edition marked the first time the entire process—from **test delivery to data collection**—was conducted entirely online.

This achievement resulted from a series of preparatory digital pretests in 2022 and 2023, which validated the system's robustness and informed the technical setup for the main operation. In 2025, the program successfully engaged a wide participant base, delivering assessments entirely online across all sampled schools.

Key innovations in 2025 included the **customization** of the TAO platform to meet PNEA's specific needs, with the **development and deployment of a fully integrated right-to-left (RTL) Arabic version, featuring personalized branding and an improved user interface**. Additionally, the implementation of a **rich item bank** enabled the transition to rotating booklets, allowing broader coverage and greater advancements in the assessment process.

To manage the scale and requirements of the assessment, a distributed and virtualized infrastructure was deployed. This architecture enabled dynamic resource allocation, load balancing, and improved system resilience, ensuring uninterrupted access and smooth performance during simultaneous nationwide connections.
The 2025 edition of the PNEA will contribute to promoting the performance of the national education system through efficient and data-driven evaluation of student learning outcomes.

CAED (BR)
Title: Building Assessment and Monitoring Platforms for Different Brazilian States: From Test Administration to Feedback
Presenter: Carlos Palacios

In 2024, **CAEd conducted assessments in over 20 Brazilian states**, each with its own unique context and set of challenges. To make this possible, each state was supported by a dedicated assessment and monitoring platform, through which the entire process was carried out - including student registration and allocation, access to and administration of print or digital instruments, and the publication of results accompanied by support materials. This entire process requires **a complex workflow and presents challenges in negotiating with various educational leaders at every stage, so that a standardized model can be effectively customized to meet each state's specific needs**.

NDET (NO)
Title: From Pen and Paper to Keyboard and Spell Checker: An Innovative Journey for Foreign Language Exams in Norway
Presenter: Kim Buxton

Exams for foreign languages in Norway have been conducted digitally since 2022. Norway has examinations in 50 different languages at three levels, corresponding to A1, A2 and B1 in CEFR. The **transition from paper-based to digital exams** has given us the opportunity to test and assess students' language skills in different ways than previously. In this presentation we would like to share some of our reflections on **our journey** so far, including **how and why decisions have been taken, and what dilemmas we are facing going forward**. This includes:

- the development of reading comprehension exercises which are automatically scored
- the introduction of digital writing, also in languages with complex writing systems
- the use of pre- and post-exam analysis for development of design and content

Agency for Higher Education, Adult education, Qualifications & Scholarships - AHOVOKS (BE-Flanders)
Title: Innovating Language Assessment: Digital testing at the Flemish Examination Board
Presenter: Heidi Dreesen and An Schotte

The Flemish Examination Board for Secondary Education provides all individuals the opportunity to obtain a secondary education diploma through self-study. Participants complete a largely digital examination program at a professionally equipped centre in Brussels, where more than 50,000 exams are administered annually.

Since 2012, the Board has been transitioning from paper-based to digital exams, with a continuous focus on improvement through technological innovation. E-testing enables efficient exam delivery, favouring closed-ended questions for automated scoring. Quality control relies on psychometric analysis to measure, monitor, and enhance the validity of exam items. This allows test developers to critically review and refine each item.

However, language exams involve skills that cannot be assessed using closed formats alone. Writing tasks require open responses scored by trained remote raters. Speaking and conversation skills are assessed by experienced examiners. Regular inter-assessor agreement analyses help monitor and support the performance of both raters and examiners.

Currently, language exams are divided into two components: a digital section (reading, listening, writing) and an oral section (speaking and conversation). Candidates may choose which section to complete first. Yet the introduction of new attainment targets, along with candidate profiles and logistical constraints, has made this structure unsustainable.

Starting in September 2025, exams will **align with the updated curriculum, which clusters productive and receptive skills**. Candidates will first complete the receptive skills exam. Upon passing, they can register for the productive skills exam, which will also be fully digital.

This reform includes a **complete revision of our scoring rubrics** to provide a uniform model for all skills and improve scoring consistency among raters.

The reorganisation has several advantages. Candidates receive additional time for both components. Speaking tasks will be digitally administered: candidates prepare individually at the exam centre, using a computer with access to source materials in a closed system and optional text-to-speech software. This setup enables more authentic language tasks and greater flexibility. It also opens the door for more frequent task updates and item-level analysis.

Through **advanced digitalisation, unified rubrics, comprehensive training, and robust analytical follow-up**, we aim to further enhance the validity and reliability of our language assessments.

In the future, we aim to explore how artificial intelligence could support us in evaluating complex language skills such as speaking and writing in a valid and reliable way. Our goal is to investigate how AI-assisted scoring can enhance both the efficiency and the quality of assessment, while maintaining fairness and transparency for all candidates.

CitoLab (NED) Innovation at CitoLab

Title: Supporting Educational Measurement with Edtech and AI with CitoLab

Presenters: Dr. Joost Kruis, Ingrid Williams, Patrick de Klein, Joey Stuiver, Nafsika Lachana

In this presentation, we will explore the **synergy between cutting-edge technology and educational assessment**. Over the past year, our focus has been on integrating AI and educational technologies to enhance the precision and efficiency of assessment processes. We will embark on five key projects that underscore our commitment to innovation and improvement in educational measurement. These will cover:

- our construction co-pilot prototype, **using AI for item content generation**, and its Q-insight integration for evaluating item and test properties.
- an introduction to our **QTI certified test player,** a robust platform for administering standardized assessments, enhancing both usability and reliability in assessment delivery.
- the initial version of **our monitoring tool, designed to examine human ratings for responses to open-ended items**. This tool facilitates better understanding and calibration of subjective scoring approaches, addressing nuances in human judgment.
- our research into **grading Dutch writing ability, utilizing machine learning and large language models** to make scoring predictions, this project aims to harness advanced computational techniques to support language assessment.
- evaluating **speech proficiency by merging state-of-the-art text-to-speech solutions** with a theoretically substantiated prediction model. As an ongoing research project, it highlights the potential of combining AI with educational theories to enhance spoken language assessments.

Through these initiatives, CitoLab strives to redefine educational measurement, emphasizing the role of advanced technology in creating meaningful, accurate, and fair evaluations that support both educators and learners.

DISCUSSION & DEMO: The FLIP+ Library and Working Groups

Title: The development of the FLIP+ Library – where are we now and where are we going?

Team: Amina Afif, Thierry Rocher, Charlotte Gill-Sotty, François-Xavier Cannes & members of the FLIP+ Working Groups

Be part of the FLIP+ journey as we unveil **the developments of the FLIP+ Library -** our digital platform designed to support international collaboration in assessment development.

Led by the FLIP+ teams, this session will include **a demo of the platform by FLIP+ DEV** as it stands today, showcasing how member institutions will be able to contribute content, navigate tools, and benefit from the Library's evolving features.

The other five **FLIP+ Working Groups - Maths, English, Psychometrics, Process Data, and Accessibility and Inclusion** - will join by sharing how they envision contributing to the Library's growth and future use. They will offer fresh insights into how content, metadata, accessibility standards, and psychometric approaches can be shaped collaboratively.

DAY 2: THEMED PRESENTATIONS: IMPLEMENTATION, PROCESS DATA & AI

DIGITAL ASSESSMENT & IMPLEMENTATION

CUPA (UK)

Title: Are Schools Ready for Digital Exams? – Using a framework to evaluate international digital readiness

Presenter: Sanjay Mistry

As education increasingly embraces technology, the **readiness of schools worldwide to transition to digital exams** remains a critical question. In this talk, we share our work in this area through the **development of a digital readiness framework** and unveil key findings from a comprehensive global study. Discover how schools are **navigating challenges such as digital infrastructure, teacher preparedness, and equity in access**. Join us to explore actionable insights, regional trends, and innovative strategies that are shaping the future of digital assessment for Cambridge schools.

NDET (NO)

Title: Digital high-stakes English exams in Norway – what have we gained?

Presenter: Kevin Steinman

While we may be well aware of the **challenges of constructing a successful digital high-stakes exam**, it is helpful to also recall the benefits. This discussion reflects on the positive aspects of digitizing summative assessment, highlighting some significant improvements that are made possible by meeting technology head-on. With a focus on English assessment in upper secondary in Norway, I will share reflections on early data highlighting improved **construct validity, accommodations, methodology, fairness and reliability**.

VRETTA (CA)

Title: quality audit framework for technology-based assessment production

Presenter: Vali Huseyn

As members of the FLIP+ community committed to high-quality, equitable, and secure assessment practices in an increasingly digital landscape, this presentation introduces a pilot project to test a structured quality audit framework for technology-based assessment production. Grounded in the **IAEA International Standards for Educational Assessment Organisations**, the proposed implementation model supports institutions in evaluating, strengthening, and standardizing their digital assessment systems, while fostering a culture of transparency and continuous improvement.

The framework covers all **critical phases of digital assessment production**:

- Design and Development, ensuring validity and fairness through secure item banking and authoring platforms
- Infrastructure and Data Security, addressing resilience, privacy, and scalability
- Human Capacity and Roles, assessing team structures and partnerships
- Scoring, Reporting, and Feedback, ensuring transparent and robust processes
- Governance and Improvement, embedding peer review and feedback cycles for continuous learning

The pilot methodology follows the IAEA's structured self-evaluation process, incorporating scenario-based self-assessment tools, interviews, process observation, gap analysis, and collaborative reflection. These methods support institutions in benchmarking their systems, identifying gaps, and planning for sustainable improvements.

Currently being piloted at Vretta, the model also explores **how global edtech providers can align with international quality standards**. Preliminary findings underscore the value of structured benchmarking in strengthening digital assessment readiness and innovation.

By sharing this pilot with the FLIP+ network, the session aims to inspire collaborative learning and build capacity for secure, effective digital assessment across diverse contexts.

DIGITAL ASSESSMENT & PROCESS DATA

Cambridge University Press and Assessment - CUPA (UK)
Title: Supporting Students' Reasoning – Developing A Digital Annotations Toolkit for High-stakes Exams
Presenter: Abdullah Ali Khan

Our research into paper-based exams reveal students do not just passively read questions before formulating responses. They underline key information in the stimulus, cross out unlikely options in multiple choice questions, mark up images like graphs to create connections between elements, and **interact with the examination using a variety of other annotation techniques**. What should this look like **in a digital environment**? In this session you will learn how we are synthesising insights from assessment research, rapid prototyping, and UX research to **create a digital annotations toolkit that supports students' thinking**.

Directorate for Evaluation, Foresight and Performance - DEPP (FR)
Title: A case study using Process Data to enhance the Interpretation of Assessment Outcomes
Presenter: Aurélie Lacroix, Franck Salles

The shift to computer-based assessments in large-scale international studies has created new opportunities to explore how students engage with mathematical problem-solving.

This presentation explores how process data—the digital traces students leave while interacting with computer-based assessments—can provide richer insight into mathematical reasoning. DEPP presents a case study from eTIMSS 2019, where middle school level students tackled an arithmetic sequence task. A multi-disciplinary framework combining cognitive theory, statistical analysis, and educational technology was used to interpret students' strategies.

Findings reveal clear links between specific behaviours and success, helping identify distinct reasoning profiles. These insights inform teacher training, formative feedback design, and the use of process data to complement traditional outcomes. Ultimately, the study shows how capturing how students solve problems can strengthen both assessment interpretation and pedagogical practice.

AQA Global Assessment Services (UK)
Title of presentation: A practical view of modularising the end-to-end process of e-assessment development and delivery
Presenter: Shaun Crowley and Simon Trevers

When planning **the implementation of large-scale digital assessments**, the conventional approach is to work with a single provider in which authoring, administration and marking is managed in the same system. However in the area of high-stakes (and traditionally paper and pen) assessments, a more flexible modular approach may be required in which best-of-breed services are selected based on how well suited they are in supporting each step in the process. Can this modular approach also apply to the **implementation of on-screen exams**? How would using separate authoring, delivery and marking systems work in practice? And what are the benefits and challenges of adopting a modular ecosystem over a single system?

In this session, Shaun Crowley from AQA Global Assessment Services will summarise the modular approach that AQA is ultimately working towards for the roll-out of onscreen GCSE exams. A core part of the session will be dedicated to demonstrating **how separate but integrated authoring and delivery systems can be used together**. As an example the presenters will show how onscreen exam content developed in a specialist authoring system (GradeMaker) can be previewed and transferred to a specialist test delivery system (Inspera) and how the marking of open responses sent from the delivery system can be managed by a specialist marking system (e-Marker®). The session will look at the practical tasks of test developers, administrators and markers. It will be followed by an exploration of the opportunities and challenges of a modular approach including a discussion around practical issues to address if considering adopting multiple integrated software products.

AI FOR BETTER DIGITAL ASSESSMENTS

CITO (NE)

Title: SchrijfBlik: Safeguarding the validity of writing assessment in the age of AI
Presenter: Tijmen Poell & Marica Balk

The rise of generative artificial intelligence (GenAI) in education necessitates **new methods for assessing student writing**. Teachers struggle to distinguish original work from AI-generated content, which challenges the validity of assessments. Only assessing the final product is no longer a fair measurement of writing skills, but banning GenAI from the classroom isn't the solution. This presentation introduces **'SchrijfBlik'**, a prototype designed to **help secondary school teachers and students in the Netherlands gain insight into the use of GenAI for writing assignments**. The aim is to make the writing process more transparent to address the challenge of safeguarding validity in the age of AI.

CAED (BRA)

Title: AI-Driven NLP Analysis of Constructed Response Items in Teacher Questionnaires: A Pilot Study
Presenter: César Soares, João Freire, Luiz Ribeiro

This study aimed to assess **levels of pedagogical reflection in teacher questionnaires** using AI-driven Natural Language Processing (NLP), capable of interpreting, analysing, and generating human language using computational techniques. Open-ended responses were pre-processed, vectorized, and analyzed using **Large Language Models (LLMs)** and a fine-tuned Portuguese language model (BERT) for accurate classification. The methodology included manual validation, synthetic data generation, and iterative training. Further analyses (n-gram frequency, Jaccard similarity, and Latent Dirichlet Allocation) investigated a prevalence of active learning approaches and varying use of assessment information. Results highlight distinct **reflective levels and instructional strategies, offering actionable insights to help build CAEd's teacher assessment program**, forwarding our mission to improve education in Brazil by advancing assessment through innovative methodologies.

AQA (UK)

Title: New Age for Digital Assessment: Explainable Scoring with Large Language Models.
Presenter: Jiazheng Li, PhD student Kings College London, Cesare Aloisi, AQA

Since 2023, there have been several studies showcasing **the potential of Large Language Models (LLMs) to be used in the scoring (marking) of open-text responses in a range of subjects. When it comes to high-stakes assessment and exams,** any use of LLMs in such fashion would require the AI systems to clearly explain the rationale behind every score awarded. This may be for legal and regulatory reasons, but also because transparent explanations build trust in the assessment process by providing accessible evidence for defending decisions when outcomes are challenged.
This presentation explores the emerging frontier of explainable automated assessment in education through the innovative use of large language models. Our research is driven by the core question: **How can we leverage LLMs not only to score student answers accurately but also to explain the scores by generating transparent, faithful rationales that mirror human rationales?**
To address this challenge, we adopted a multi-staged approach. First, we **distilled ChatGPT's reasoning process** into smaller, local LLMs to assess student responses and explain scoring decisions. Following that, we optimised our local LLMs to generate more accurate and faithful scoring explanations by **leveraging recent human preference alignment technology to optimize LLMs** in generating more accurate and faithful assessment explanations to improve reliability. We then developed a dual-model reasoning framework that turns marking into an automated proposal-critique-refinement process in a way that simulates human debates. Finally, we **developed an interactive system, AERA Chat,** to provide educators with a visual and analytical platform to interact with our AI models for real-time assessment and rationale verification. Preliminary results are promising, showing notable improvements in scoring accuracy and rationale quality compared to traditional methods. However, challenges remain, including the occasional generation of incorrect assessment rationales and limited generalizability to unseen questions. In the talk, we will discuss these limitations and outline the next steps for developing more robust, scalable, and transparent automated assessment systems that can better support educators and learners in the digital age.

PRESENTERS, PANELISTS AND CONTRIBUTORS

Amina Afif holds Master's degrees in both Statistics and Education. With over 15 years in Luxembourg's Ministry of Education, she led the division for school quality development and data use, representing the country in international education policy forums. As an experienced project leader, she has coordinated multilingual teams across national and international initiatives, working with policymakers, researchers and practitioners in the field of educational policy, evaluation, digital assessment, data use and school improvement. Since 2022, she leads professional development programmes for school leaders and school boards in Seychelles, supporting the national school autonomy initiative. Amina now works independently as an education policy advisor and transformational life coach. She is also a founding member and Executive Secretary of the FLIP+ e-assessment association and serves as Project Executive for the international FLIP+ Library initiative.

Cesare Aloisi is Head of AI for Assessment Innovation at AQA – the most chosen general qualifications exam board in England. The team focuses on how to use AI to improve assessment for teachers and learners – from question authoring to marking of exams and classroom assessments.

Cesare's current areas of interest are around machine appraisal of human performance (e.g. can an AI identify competent human writers, mathematicians, dancers…?) and, relatedly, human appraisal of machine performance (e.g. how do we find out reliably and systematically what an AI knows and can do?).

Marica Balk, MSc. is an innovation expert with a background as concept developer, trend researcher and assessment specialist. She is a pioneer in developing digital assessment tools and translates complex issues into applicable products. She has been involved in creating a variety of assessment tools. Currently she is exploring how AI can enhance assessments by developing various concepts like 'SchrijfBlik'. Additionally, she is project manager of a granted large-scale survey on reading skills (2023-2026). Her aim is to contribute to a pleasant assessment experience and appropriate decisions by putting the user first.

Kim Buxton is a Senior Adviser at the Norwegian Directorate for Education and Training. He primarily works with assessment and curriculum administration for foreign languages and minority languages in Norway. Prior to joining the Directorate in 2010, he taught German and Italian at both lower and upper secondary levels.

Alberto Díaz-Roncero Canales: Technical Teaching Advisor responsible for the design and rollout of large-scale diagnostic assessments in Spain. Educational psychologist and teacher with experience in multicultural and technological environments.

Shaun Crowley, Director of International Business Development, AQA Global Assessment Services. Shaun is part of the AQA Global team, an AQA subsidiary responsible for supporting governments and assessment providers with services including technology and consultancy from AlphaPlus. Shaun is responsible for developing assessment solutions for international governments and was previously in a business development role for GradeMaker prior to its acquisition by AQA. Shaun has a background in international qualifications and educational publishing.

Patrick de Klein is a UX/UI Designer and Developer at Cito, where he works on developing features for the open-source QTI 3.0 player. He is passionate about creating digital experiences that are both intuitive and inclusive, ensuring they are accessible to all users. Together with his team, he creates innovative assessment tools through prototyping for the education of tomorrow. Driven by the belief that QTI should be easier to implement, they started developing their own QTI player as a side project. This open-source library was recently awarded 1EdTech certification and received recognition for its strong focus on interoperability.

Heidi Dreesen studied Dutch Linguistics and Literature at the Vrije Universiteit Brussel (VUB). She gained experience as a Dutch teacher, subject coordinator & equal educational opportunities coordinator in a secondary school in Louvain. In 2019, she joined the Flemish Examination Board as an examiner and test developer. She later became part of the data analysis team and currently focuses on developing a new assessment tool.

João Freire: Educational Assessment Specialist at CAEd, MsC in Applied Economics and PhD candidate in Computational Modeling. Currently working with machine learning applications in educational assessment, namely the automated prediction of IRT parameters for new items, as well as item generation with LLMs.

Hugo Hopstaken is programme manager at the Dutch Board for Tests and Examinations (CvTE), currently leading the national programme for the digitalisation of central exams. The programme focuses on developing a future-proof digital

examination system, addressing applied skills assessment, accessibility, and logistical improvements. Prior to his current role, he led various educational projects and programmes across the dutch education system.

Vali Huseyn is an educational assessment expert, with significant contribution to modernizing assessment practices across government and private sectors. At The State Examination Centre of Azerbaijan, he was instrumental in enhancing assessments and leading initiatives such as developing a unified testing platform and promoting assessment literacy. At Vretta, Vali leads strategic assessment initiatives, aligning them with organizational goals and educational standards, and drives innovative partnerships to boost assessment effectiveness and educational results.

Noura Kelloul is in charge of IT development within the PNEA team. State Engineer specialized in Computer Science, Networks, and Systems, graduated from the National Institute of Posts and Telecommunications (Rabat, Morocco).

Abdullah Khan is a Digital Assessment Lead at Cambridge University Press & Assessment. In this role, he supports the development of assessment models for digital products and qualifications by integrating insights from academic and user research. His work includes iteratively refining digital item types to enhance construct validity as well as candidate experience, managing the development of a digital annotations toolkit, and exploring comparability issues arising from launching digital exams.

Before joining Cambridge, Abdullah worked across nonprofit, policy research, and edtech sectors in Pakistan, focusing on evidence-based, culturally responsive support for teachers. At The Citizens Foundation, he led a team designing virtual training for 800+ school principals. As an EdTech Hub consultant, he collaborated with Bangladesh's Aspire 2 Innovate unit to enhance teacher-led evidence generation. Abdullah is a Fulbright scholar with a Master's degree in Mind, Brain, and Education from the Harvard Graduate School of Education. He started his career as a teaching fellow with Teach For Pakistan.

Dr. Joost Kruis received his Ph.D. in Item Response Mechanics from the Department of Psychological Methods at the University of Amsterdam. As a psychometric researcher he heads the AI research programme at Cito, where his current research focuses on applications of ML and AI in education. Additionally, he is involved in the norming and equating of the annual Dutch national secondary school exams.

Aurélie Lacroix leads the statistical unit for secondary national assessments at the Directorate for Evaluation, Foresight and Performance (DEPP) within the French Ministry of National Education. She works on the analysis of large-scale educational assessments, both national and international, to support data-informed educational policy and reporting. She is also a member of the psychometric working group of Flip+. Her work helps shed light on student achievement and supports the ongoing development of the French education system.

Nafsika Lachana is completing her Master's in Artificial Intelligence at Utrecht University. Her thesis, conducted in collaboration with Cito, focuses on automated scoring systems for open-ended questions in Dutch. Her research explores effective approaches for processing Dutch-language responses and developing interpretable models for multi-aspect scoring aligned with educational rubrics. She holds a Bachelor's degree in Mathematics from the University of Patras, Greece.

Jiazheng Li is a third-year PhD candidate at King's College London, UK, where he develops explainable automated student scoring systems under the supervision of Prof. Yulan He, with funding from AQA Education. Prior to his doctoral studies, he earned a First-Class Honours Bachelor's degree in Computer Science from University College Dublin in Ireland. His research interests include machine learning, large language models, and explainable AI. He has published extensively and actively contributed to the research community as a reviewer and chair at top NLP/AI conferences such as EMNLP, ACL, and AAAI.

Laura Verdasco Menendez is a research fellow at AQA. As part of the Psychometrics and Innovation team, she uses psychometrics to evaluate assessment-related data. Recently, her focus includes the use of item-response theory to improve the maintenance of assessment standards, and evaluating novel approaches to mark estimation.

Sanjay Mistry is Head of Digital Insight and Impact at Cambridge University Press and Assessment. The role involves leading on the definition and delivery of the research and insight strategy that is pivotal to underpinning the organisation's strategic roadmap for the development of the digital products and services portfolio that serve the international and UK markets, and working with schools, school groups and Associates to bring them on Cambridge's digital journey. Sanjay had a successful career in Primary education and leadership for 11 years before joining Cambridge International in 2012.

Sanjay has a passion for really understanding the affordances that technology brings to improving the educational outcomes for all learners and ensuring it is used to add value to the overall teaching, learning and assessment experience. As well as being a subject matter expert on the digital high stake's exams programme, he also leads on projects around digital innovation, such as the approach to the accessibility of digital exams, understanding global markets readiness for digital exams, assessment development in the Early Years and more recently, the value that Artificial Intelligence can bring to digital exam journey.

Tijmen Poell is a member of the Citolab prototypes team. With a background in Industrial Design and a focus on UX/UI, he brings a design-driven approach to educational technology. He has contributed to various innovative projects, including the design and development of 'SchrijfBlik'. His firsthand experience as a teacher in secondary education, enables him to create solutions that resonate with both educators and students.

Hannah Rowe is the Deputy Director of Assessment at AlphaPlus, working in the design and delivery of a broad range of different assessments - from managing the content delivery of a national scale assessment for primary and early secondary students, to the design of apprenticeships and professional qualifications for adults. Before to joining AlphaPlus Hannah worked at AQA in assessment design and prior to this, she was a teacher and senior manager in secondary and sixth form education. Hannah has recently completed a Masters in Education at the University of Cambridge, where her dissertation focused on teacher assessment literacy and its role in formative assessment.

Carlos Palacios is a Research and New Projects Supervisor at CAEd. He has a PhD in Language at the Federal University of Rio de Janeiro (UFRJ) and conducted post-doctoral research on public policies for literacy at the Department of Education at the Pontifical Catholic University of Rio de Janeiro (PUC-Rio). At CAEd/UFJF, he leads a team focused on research, innovation and partnerships with international institutions in large-scale assessment.

Elodie Persem is a former teacher of literature specializing in working with disabled students. She has trained teachers and trainers for many years on the issues of school inclusion, managing heterogeneity and teaching cross-curricular skills. Since joining DEPP (Directorate for Evaluation, French Ministry of Education), she has been put in charge of the "AIR" unit (Accessibility, Innovation and Research) as a specialist in inclusive education. This department was created to improve the accessibility of assessments and to develop partnerships with researchers in order to build assessments around the so-called 21st skills.

Luiz Ribeiro: Educational Assessment Specialist at CAEd, sociologist with a PhD in Social Sciences (UFJF, Brazil). Applies psychometric techniques to enhance educational evaluation and investigates factors associated with student performance through statistical modeling.

Franck Salles is a research and evaluation officer at the Directorate for Evaluation, Foresight and Performance (DEPP) within the French Ministry of National Education. He has a background in mathematics education. His work focuses on national and international assessments, particularly in mathematics. He contributes to design assessment content, implement data collection and produce reports on education in France and plays a key role in managing large-scale studies such as PISA and TIMSS. A regular contributor to official DEPP publications, he is also coordinating the group for international assessments at DEPP.

An Schotte studied Germanic Philology (Dutch/English) at Ghent University. She worked in various educational settings in Ghent (secondary and adult education) until 2014. That year, she joined the Flemish Examination Board as a test developer and examiner. She currently leads the Dutch subject group and coordinates the language department.

César Soares: Educational Assessment Specialist at CAEd, psychologist by training, with a Ph.D. from the University of São Paulo, Brazil, complemented by postdoctoral research at the University of Saskatchewan, Canada. My current work focuses on applying Natural Language Processing methodologies and psychometric techniques to advance educational evaluation practices.

Kevin Steinman is a Senior adviser at the Norwegian Directorate for Education and Training. His five years there have coincided directly with Norway's digitization of high-stakes exams after curricular renewal. So he may be slightly biased when it comes to whether or not to digitize. He comes from a long line of teachers, and when not at work, enjoys baseball (both playing and watching), tennis, cycling and learning to swim freestyle.

Joey Stuiver is a psychometric researcher at Cito, who focuses on developing tools to enhance educational assessment. With a Master's degree in Cognitive Neuroscience and a bachelor's minor in psychology, he brings a unique perspective, combining insights from different fields. Although new to the field of psychometrics, he has discovered a passion for data science, coding and development, which have become integral to his work.

Thierry Rocher is the Director of Student Evaluation at the Office for Student Assessment (DEPP) at the Ministry of Education in France. He is also the Chair of the IEA and the President and Founding member of the FLIP+ e-assessment association.

Simon Trevers, Director of Assessment Strategy and Propositions, Inspera. Simon is responsible for developing Inspera's strategy and propositions, and works closely with awarding bodies to ensure their needs are understood and met. With over 18 years experience in EdTech and Digital Assessment, Simon has worked with many of the largest awarding organisations and professional certification bodies, designing and implementing digital assessment services that deliver engaging, accessible assessments in a robust, reliable and scalable manner.

Sabrina Gallego Verdi: Technical Teaching Advisor responsible for the design of multilingual items for diagnostic assessments in Spain. Technical Teaching Advisor responsible for the design of multilingual items for diagnostic assessments in Spain. EFL educator with experience in multilingual and international education and recurrent trainer at universities and professional development centres.

Ingrid Williams is a Research and Design Specialist at CitoLab, a multidisciplinary innovation team with a focus on the development of prototypes for educational measurement. As an educational innovator and assessment expert, Ingrid explores how to make future assessments more engaging, accessible and effective using design research.