

AIME-CON 2026



TABLE OF CONTENTS

Schedule at a Glance	8
Floor plan	9
Keynote: Alina A. von Davier.	13
Keynote: Dragan Gašević	14
Monday, October 5	15
Tuesday, October 06	22
Wednesday, October 07	61

Access the most up-to-date schedule
via the online program.



DOWNLOAD THE APP!



SEARCH IN STORE:
eConference.io

Download, install and open the eConference.ioapp.
AIME26



This is an event app for attendees of the 2026 Artificial Intelligence in Measurement and Education Conference, presented by the National Council on Measurement in Education, taking place in Pittsburgh, October 5–7, 2026.





AIME-Con 2026

October 5–7 | Pittsburgh

This year's conference theme, "Measurement Science in AI-Integrated Assessment and Pedagogy," explores the vital intersection of AI innovation and psychometric rigor, ensuring that the next generation of educational tools is anchored in validity, reliability, and fairness.

AIME-Con brings together an interdisciplinary group of leading experts in AI, psychometrics, education, NLP, and learning analytics to bridge gaps between these fields.

THANK YOU TO OUR SPONSORS

PLATINUM



***ets research institute**

GOLD

Curriculum Associates



SILVER



THANK YOU TO OUR SPONSORS

SUPPORTER



The National Council on Measurement in Education (NCME)

The National Council on Measurement in Education is a community of measurement scientists and practitioners who work together to advance theory and applications of educational measurement to benefit society. A professional organization for individuals involved in assessment, evaluation, testing, and other aspects of educational measurement, our members are involved in the construction and use of standardized tests; new forms of assessment, including performance-based assessment; program design; and program evaluation. Learn more about NCME, including our goals and our leadership, at www.ncme.org.

NCME Special Interest Group on Artificial Intelligence in Measurement and Education (AIME)

This special interest group (SIGIMIE) seeks to advance the theoretical and applied research into AI of educational measurement by bringing together data scientists, psychometricians, education researchers, and other interested stakeholders. The SIGIMIE will discuss current practices in using Generative AI, approaches to evaluate their precision/accuracy, and areas where more foundational research is required into the way we test and measure educational outcomes. This group seeks to create a strong professional identity and intellectual home for those interested in the use of AI in many areas, including automated scoring, item evaluation, validity studies, formative feedback, and generative AI for automated item generation.

Proposal Requirements and Review Process for Full Papers

AIME-Con invited submission of "Full Papers," which were submissions of up to six pages (excluding references, tables, and figures), prepared using the ACL LaTeX or Word templates. These papers presented completed research or theoretical work intended for inclusion in the published conference proceedings. Submissions included a title (≤ 12 words), a brief abstract (≤ 50 words), a designated topic of interest, and the full paper. Submissions were blinded for peer review. Submissions were evaluated by members of the review committee using a rubric that evaluated the following dimensions:

- **Relevance and community impact:** pertinence to the AI in measurement and education community, and potential contribution to current discussions and challenges in the field.
- **Significance and value:** scholarly merit or practical importance of the work, and potential impact on theory, practice, or policy.
- **Methodological rigor:** coherence and appropriateness of the proposed methods, techniques and approaches; and soundness of the overall research design.
- **Quality of expected outcomes:** whether the proposed analysis and interpretation methods are appropriate, and the potential contribution to knowledge in the field.
- **Feasibility and timeline:** the realistic likelihood that the proposed work can be completed by the conference date.

Both structured summaries and full papers were assessed on equal footing, with expectations tailored to each format. For the purposes of this conference, "AI" was defined broadly to include rulebased methods, machine learning, natural language processing, and generative AI/large language models. Reviewers provided constructive feedback and overall recommendations to ensure that accepted sessions reflected both scholarly merit and practical value to the AI in measurement and education community.

Proposal Requirements and Review Process for Work-in-Progress Papers

AIME-Con invited submissions of Working Papers, which were submissions of up to 1,000 words in the form of a structured summary. This format was designed for work-in-progress or pilot studies. Working Papers required a title, short abstract, and followed a structured format with the following headings:

- Background
- Aims
- Sample(s)
- Methods
- Results (or Anticipated Results)
- Conclusions (or Anticipated Contributions)
- References, tables, and figures included as needed.

Submissions were evaluated by members of the review committee using a rubric that evaluated the following dimensions:

- **Relevance and community impact:** pertinence to the AI in measurement and education community, and potential contribution to current discussions and challenges in the field.
- **Significance and value:** scholarly merit or practical importance of the work, and potential impact on theory, practice, or policy.
- **Methodological rigor:** coherence and appropriateness of the proposed methods, techniques, and approaches; and soundness of the overall research design.
- **Quality of expected outcomes:** whether the proposed analysis and interpretation methods are appropriate, and the potential contribution to knowledge in the field.
- **Feasibility and timeline:** the realistic likelihood that the proposed work can be completed by the conference date.

Both structured summaries and full papers were assessed on equal footing, with expectations tailored to each format. For the purposes of this conference, "AI" was defined broadly to include rulebased methods, machine learning, natural language processing, and generative AI/large language models. Reviewers provided constructive feedback and overall recommendations to ensure that accepted sessions reflected both scholarly merit and practical value to the AI in measurement and education community.

Proposal Requirements and Review Process for Coordinated Paper Sessions

AIME-Con invited submissions of coordinated paper sessions, which brought together 4-5 papers on a common theme within a 90-minute session. Each proposal included both session-level information (title, abstract, keywords, chair/moderator, and discussant where applicable) and paper-level details (title, short abstract, topic of interest, and either a 1,000-word structured summary or a sixpage paper). All contributors were identified at submission, as the review process was not blind. Submissions were evaluated by members of the review committee using a rubric that evaluated the following dimensions:

- **Relevance and community impact:** pertinence to the AI in measurement and education community, and potential contribution to current discussions and challenges in the field.
- **Significance and value:** scholarly merit or practical importance of the work, and potential impact on theory, practice, or policy.
- **Methodological rigor:** coherence and appropriateness of the proposed methods, techniques, and approaches; and soundness of the overall research design.
- **Quality of expected outcomes:** whether the proposed analysis and interpretation methods are appropriate, and the potential contribution to knowledge in the field.
- **Feasibility and timeline:** the realistic likelihood that the proposed work can be completed by the conference date.

NCME Leadership

Kadriye Ercikan, Ph.D., President
Susan Lyons, Ph.D., Executive Director

AIME-Con Conference Chairs & Program Committee

Joshua Wilson, Ph.D., University of Delaware
Christopher Ormerod, Ph.D., College Board
Magdalen Beiting Parrish, Ph.D., Federation of American Scientists

AIME-Con Program Committee

Polina Harik, Ph.D., NBME

SCHEDULE AT A GLANCE

MONDAY, OCTOBER 5

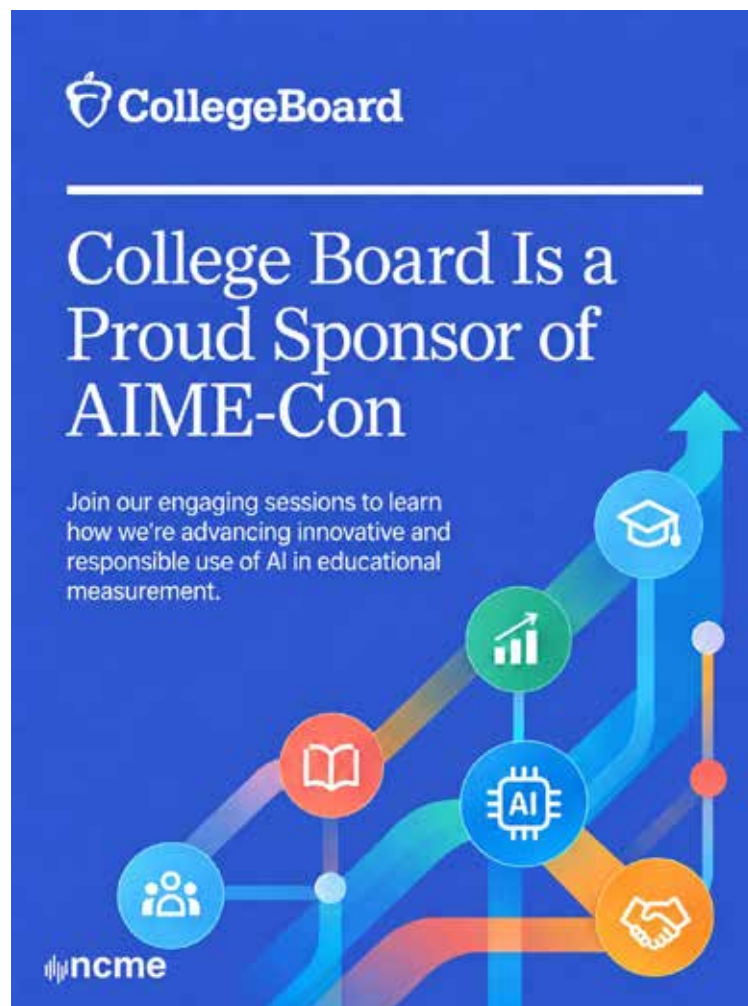
7:30 a.m.–4:30 p.m.	Conference Check-In/Info Desk Open
9 a.m.–11 a.m.	Two-Hour Trainings (optional add-ons during registration)
1 p.m.–5 p.m.	Four-Hour Trainings (optional add-ons during registration)

TUESDAY, OCTOBER 6

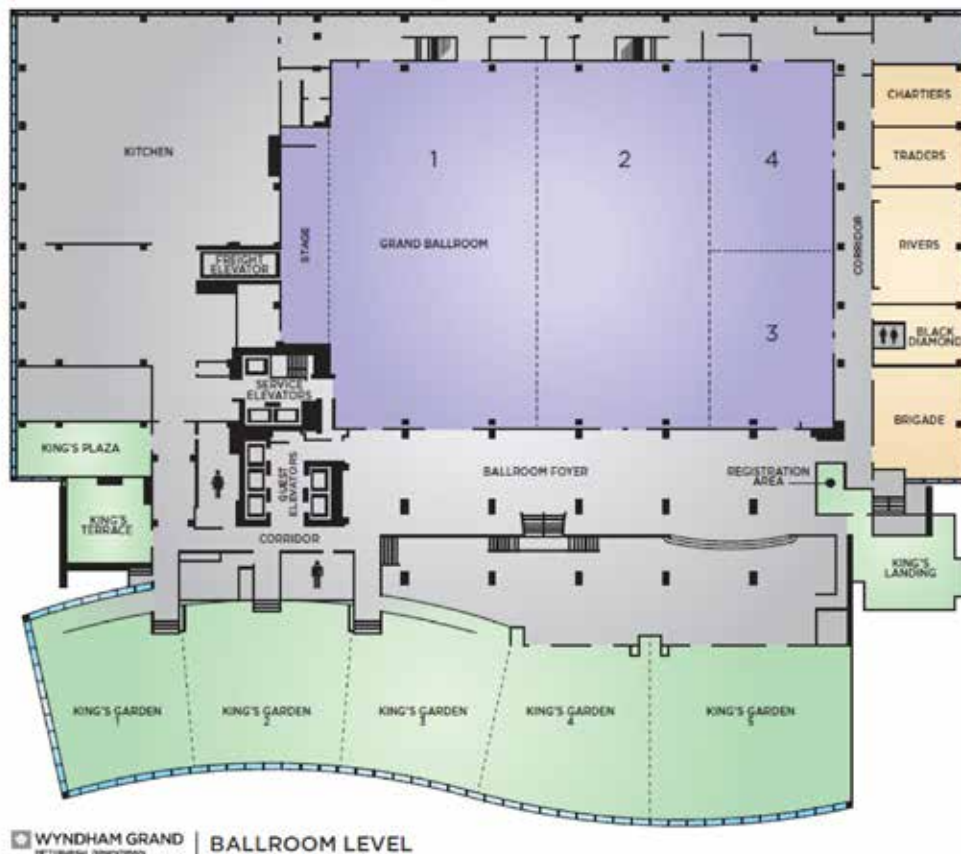
7 a.m.–5 p.m.	Conference Check-In/Info Desk Open
7 a.m.–8 a.m.	Breakfast (provided)
8 a.m.–9 a.m.	Keynote Session
9 a.m.–9:15 a.m.	Break
9:15 a.m.–10:45 a.m.	Session Block 1
10:45 a.m.–11 a.m.	Break
11 a.m.–12:30 p.m.	Session Block 2
12:30 p.m.–1:30 p.m.	Lunch (provided)
1:30–1:45 p.m.	Break
1:45 p.m.–3:15 p.m.	Session Block 3
3:15 p.m.–3:45 p.m.	Afternoon Beverage Break
3:45 p.m.–5:15 p.m.	Session Block 4
5:15 p.m.–6:30 p.m.	Reception with Posters

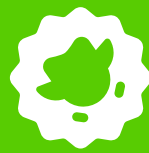
WEDNESDAY, OCTOBER 7

7 a.m.–4 p.m.	Conference Check-In/Info Desk Open
7 a.m.–8 a.m.	Breakfast (provided)
8 a.m.–9 a.m.	Keynote Session
9:00 a.m.–9:15 a.m.	Break
9:15 a.m.–10:45 a.m.	Session Block 5
10:45 a.m.–11:00 a.m.	Break
11:00 a.m.–12:30 p.m.	Session Block 6
12:30 p.m.–1:30 p.m.	Lunch (provided)
1:30–1:45 p.m.	Break
1:45 p.m.–3:15 p.m.	Session Block 7
3:15 p.m.–3:45 p.m.	Afternoon Beverage Break
3:45 p.m.–5:15 p.m.	Session Block 8
5:15 PM–6:30 PM	Reception



FLOOR PLAN

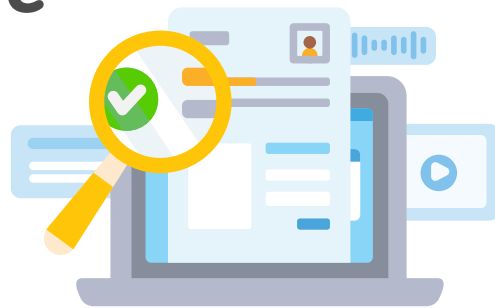




duolingo english test

The future of language assessment is here

The Duolingo English Test is a computer adaptive test powered human-in-the-loop AI and supported by rigorous validity research. The test measures speaking, writing, reading, and listening skills, providing a deeper insight into English proficiency.



Built on the latest language assessment science

- ✓ Accessible by design, supporting test takers wherever they are for just \$70
- ✓ Built on rigorous research and industry- leading security
- ✓ Integrates the latest assessment science and AI for accurate results
- ✓ Accepted by over 6,500 programs worldwide



englishtest.duolingo.com



AI research that empowers every learner

ETS Research Institute advances responsible AI research in learning, education and assessment.

We're applying decades of measurement science expertise to ensure AI-powered learning and measurement solutions are valid, equitable, fair and transparent. ETS Research Institute promotes the responsible use of AI by conducting rigorous research, creating practical tools and setting meaningful standards.



Learn more about our research on AI for assessment capabilities



Discover our AI research expertise at NCME AIME 2026:

- Validating AI-simulated students to draw conclusions about teacher practice
- Testing AI-generated reply consistency across model changes
- Promoting individual-level fairness in automated scoring — and reporting the accuracy tradeoff
- Verifying that AI-personalized math problems remain mathematically equivalent to standard items
- Documenting how AI-generated reading items can match statistics yet require different inferences

The AAMC Strengthens the World's Most Advanced Medical Care



Representing institutions that deliver the world's most advanced medical care, the AAMC provides programs and services that support the entire spectrum of education, research, and health care. We believe in medical education that prepares physicians and scientists to meet the nation's evolving health needs, support a diverse and culturally competent health care workforce, lead medical advancements that prevent disease and alleviate suffering, and more.

Learn more at aamc.org.

26-050 (01/26)

AAMC.ORG

Driving reliable **impactful assessment** change in education through and data-driven insights.



**Advancing
the Standard**
View our
Research





Measure what matters

One connected learning story

Navy+ brings skills-based MTSS data and standards-level classroom insight together in one platform, helping educators see where students are, what they need, and how to respond with greater confidence.

By connecting screening, progress monitoring, and classroom formative data, Navy+ makes it easier to spot patterns, personalize support, and follow growth over time, without jumping between systems or piecing together disconnected reports.

Visit PearsonAssessments.com/navvy to learn more

800-627-7271 | PearsonAssessments.com

© 2026 Pearson Education, Inc. or its affiliates. All rights reserved. Pearson is a trademark, in the US and/or other countries, of Pearson plc

KEYNOTE

Alina A. von Davier

Alina A. von Davier is a researcher, innovator, and an executive leader with over 20 years of experience in EdTech and in the assessment industry. She is currently the Chief of Assessment at [Duolingo](#), where she leads the [Duolingo English Test](#) research and development area. She is also the Founder and CEO of [EdAstra Tech](#), a service-oriented EdTech company. In 2022, she joined the [University of Oxford](#) as an Honorary Research Fellow, and [Carnegie Mellon University](#) as a Senior Research Fellow. She currently serves as a Non-executive Director on the Board for [MACAT](#), an EdTech company focused on critical and creative thinking, learning and assessment and she is a Venture Partner for [LearnLaunch](#) Fund and Accelerator. Her research is in the field of computational psychometrics, machine learning, and education.



Von Davier was a Chief Officer at ACT, where she led ACTNext, an innovation unit. Previously she was a Senior Research Center Director at Educational Testing Service. Von Davier earned an M.S. in Mathematics from the [University of Bucharest](#) in 1990, and a Doctorate in Mathematics from [Otto von Guericke University Magdeburg](#) in 2000. She also completed classes in an Executive MBA program from Harvard Business School in 2019.

Von Davier's work has been widely recognized in the academic community. In 2019, she was a finalist for the Innovator award from the EdTech Digest. In 2020, she received ATP's Career Award for her contributions to assessment. The American Educational Research Association awarded her the Division D Signification Contribution Educational Measurement and Research Methodology Award for her publications "Computerized Multistage Testing: Theory and Applications" (2014) and an edited volume on test equating, "Statistical Models for Test Equating, Scaling, and Linking" (2011). In 2022 she also received the Brad Hanson award from National Council in Education (NCME) for her work on adaptive testing and for the co-authored book on computer adaptive testing with R.

See von Davier's GoogleScholar page:

<https://scholar.google.com.br/citations?user=Eu9DtEsAAAAJ&hl=en>



ADVANCING ASSESSMENT, SUPPORTING OPTIMAL CARE

Through research and collaboration, NBME is evolving how we evaluate and support learners, with a focus on applying new technology to develop assessments that measure and build the knowledge and skills needed to provide optimal, effective care to all.

innovationsinassessment.org



KEYNOTE

Dragan Gašević

Professor Dragan Gašević is a leading scholar in learning analytics and educational data science, currently based at Monash University, where he serves as a Professor and Director of the Centre for Learning Analytics. His work focuses on understanding and improving learning and teaching through the analysis of large-scale educational data, with particular emphasis on self-regulated learning, assessment, and the design of effective digital learning environments.

Professor Gašević has made significant contributions to bridging the fields of education, data science, and artificial intelligence, helping shape learning analytics as a mature research discipline. He is widely published in top-tier journals and has received multiple awards for his research, including recognition from international educational and computing societies.



Beyond his research, he plays an active leadership role in the global learning analytics community, including editorial responsibilities for major journals and involvement in key conferences such as LAK (Learning Analytics & Knowledge). His work has had substantial impact on both academic research and institutional practices, informing how universities use data to enhance student success and teaching quality.

Connecting Data to Instruction



 **i-Ready Inform™**

LEARN MORE

FULL SCHEDULE

MONDAY, OCTOBER 5

Beyond Bots: Promoting and Measuring with Facilitative AI Training Session (2 hour)

9:00 AM – 11:00 AM

Conference Center: Commonwealth 1

*Presenter(s): Simon Cullen, Co-Founder & President | Disagree Wisely - Senior Fellow | Johns Hopkins School of Government & Policy; Michael Strambler, Yale School of Medicine Division of Prevention and Community Research
Yale Child Study Center Director, Child Wellbeing and Education Research Co-Director, Partnership for Early Education Research*

Abstract:

Most AI-in-measurement work casts AI as a generator or scorer of student work. This hands-on session introduces a third role—facilitative AI—in which AI structures human-to-human interaction so the interaction itself becomes measurement evidence. Through a live demonstration, participants explore novel ways to measure reasoning in real time.

Session Description: Across this conference, artificial intelligence plays two roles: it generates student-facing material (items, passages, feedback) and it scores student work (essays, constructed responses, spoken language). While both are valuable and will be well represented in our presentation, we introduce a third role that is nearly absent from the field's imagination: facilitative AI. Here, AI neither writes nor grades. It structures the cognitive environment a person engages with—and the person's interaction with that environment becomes the measurement itself.

This matters most for constructs that conventional instruments struggle to reach. Consider critical thinking. We can measure it with the Halpern Critical Thinking Assessment, the CCTST, or LSAT logical-reasoning items. But a strong score shows mainly that a student will reason carefully when they know they are being tested. The question that matters for education is whether they have acquired the disposition to reason well in everyday life, such as in real disagreements, with real people. That disposition is hard to see in a test taken alone, and it is not elicited well by a student talking to a chatbot, because authentic human-to-human exchange engages social, cognitive, and emotional capacities that a machine might not trigger. Facilitative AI can stage a genuine disagreement between two (human) students, keep it constructive, and turn the resulting interaction into rich evidence. The same logic extends to social-emotional learning, collaboration, and other constructs that live in interaction rather than on answer sheets.

Sway—an AI-facilitated discussion platform used by more than 11,000 students at over 100 colleges and universities—serves as an example of this paradigm, though it is only one early exploration of a much larger space. Participants will experience it directly in a live demo: we will match participants into pairs across a real disagreement, facilitate their discussion with Sway, and then watch a class-level measurement report built from their own conversations. Measurement domains include opinion movement, mutual understanding, discussion themes, common ground, persistent disagreements, and where the AI facilitator intervened or, tellingly, stayed silent. The session's aim is not to teach a single tool or a single method. Rather, it is to expand the conceptualization of AI's role in education toward what AI can facilitate and measure between humans and together with human reasoning.

Software Requirements

Software Needed. Sway, a free, web-based AI-facilitated discussion platform. It runs in any modern browser. We provide temporary access codes and pre-loaded discussion prompts in advance. If on-site connectivity limits live report generation, we run a pre-captured report from a real class as a seamless fallback.

Installation/Payment Requirements. None — no installation and no purchase. Sway is free, and nothing is sold or marketed during the session.

Hardware Requirements. No special hardware; participants need only a personal device — a laptop, tablet, or phone — and Wi-Fi.

FULL SCHEDULE

MONDAY, OCTOBER 5

How to generate scenario-based learning and assessment for post-secondary courses with GenAI Training Session (2 hour)

9:00 AM – 11:00 AM

Conference Center: King's Garden 3

Presenter(s): Benjamin Pierce, Institute for Intelligent Systems; Zuowei Wang, ETS; Tenaha O'Reilly; John Sabatini, Institute for Intelligent Systems, University of Memphis; Joseph Magliano, Georgia State University; Jonathan Cohen, Georgia State University; Michael Suhan, ETS

Abstract:

This interactive two-hour workshop introduces the design and use of GenAI-enabled scenario-based learning assessments (SBLAs), which place knowledge and skills in realistic contexts. Participants will co-create an SBLA together in order to learn how to address the risks of GenAI interfering with meaningful learning and assessment. Bring an internet-connected laptop.

Session Description: To address the challenges introduced by GenAI while also maximizing its potential benefits, this session focuses on creating Scenario-Based Learning Assessments (Sabatini et al., 2020). SBLAs provide a realistic goal to engage learners in authentic problem solving. SBLAs help students build knowledge, model best practices, foster critical evaluation of sources of information, and encourage application of their learning to solve authentic problems. SBLAs provide a structured environment for students to interact reflectively with GenAI during learning and thus are inherently resistant to GenAI's shortcuts.

As part of two multi-year, grant funded projects, development teams led by post-secondary instructors have been working with assessment developers, learning scientists, instructional designers, teacher trainers, and GenAI/technology developers to build:

1. a shareable library of exemplary SBLA prototypes across diverse disciplines (e.g., psychology, interdisciplinary studies, health science, writing);
2. an Open-source, GenAI-driven Authoring System for use by instructors;
3. a Professional Support System such as this training;
4. a Multimodal Media Literacy Framework and associated Wiki; and
5. evaluate the system and scale-up dissemination across post-secondary contexts.

The projects have the dual aim of preparing instructors to develop SBLAs and for those instructors to incorporate GenAI into the SBLA designs to foster effective and responsible AI literacy skills in their students. As the project completes its second year, the team has developed and iteratively piloted multiple SBLA prototypes in multiple disciplinary courses, has a working model of a GenAI authoring system, and has been conducting training workshops to hone the professional development system.

The proposed training session builds on progress so far in these projects and is structured as a 2-hour interactive workshop, designed to help post-secondary instructors and measurement specialists to create SBLAs utilizing GenAI authoring. The session begins with an overview of SBLAs, highlighting the design principles and their unique benefits for teaching, learning, and assessment. Participants will learn how to use SBLAs to support authentic learning and assessment practices while learning techniques to help counter (or at least mediate) GenAI-driven challenges.

Next, attendees will be guided in how to use GenAI to support and co-create an SBLA with an interdisciplinary facilitation team (consisting of psychometricians, learning scientists, and technical specialists), and critically discuss ways that SBLAs support student engagement with AI as well as the instructional and measurement issues raised by this engagement.

The session concludes with a reflection and open floor discussion of the whole process, which will offer participants the opportunity to address specific questions, share insights, and discuss implementation strategies within their own educational contexts.

Software Requirements

N/A

FULL SCHEDULE

MONDAY, OCTOBER 5

Generative AI for Item Development: A Hands-On Workshop Training Session (4 hour)

1:00 PM – 5:00 PM

Conference Center: Commonwealth 2

Presenter(s): Henry Makinde; Hope Adegoke, University of North Carolina, Greensboro; Mubarak Mojinyinola

Abstract:

This interactive, hands-on session introduces educators and measurement professionals to generative AI and then its application in item generation, covering prompt engineering, retrieval-augmented generation (RAG), fine-tuning, and agentic systems. Participants implement each technique live in Google Colab using large language models (LLMs), gaining practical, immediately applicable skills.

Session Description: Learning Objectives

By the end of this session, participants will be able to:

- Explain the core concepts of generative AI and large language models (LLMs) and how they apply to educational measurement.
- Apply prompt engineering strategies (zero-shot, few-shot, and chain-of-thought) to generate assessment items.
- Build a retrieval-augmented generation (RAG) pipeline for context-aware, standards-aligned item generation.
- Fine-tune an LLM and configure a simple agentic workflow for specialized item-generation tasks.
- Implement each technique in Python using cloud-based notebooks GoogleColab, requiring no local setup.

Content and Relevance

Artificial intelligence is rapidly transforming educational measurement, and generative AI in particular is reshaping how assessment items are written and reviewed. Traditionally, item writing and review are performed by subject-matter experts (SMEs) through a comprehensive development and evaluation process. This gave rise to automatic item generation (AIG), which still depends heavily on SMEs to construct cognitive models. Generative AI now lowers this barrier substantially, and a growing body of research demonstrates how LLMs can support item generation directly.

This session equips item writers, assessment experts, instructional designers, and measurement practitioners to navigate this paradigm shift. It progresses from foundational to advanced techniques: we begin with prompting strategies (zero-shot, few-shot, chain-of-thought, tree-of-thought, and others) and RAG, then lay the foundations of fine-tuning and agentic systems.

Throughout, we connect each method back to measurement priorities validity, reliability, fairness, and alignment consistent with the conference theme.

All demonstrations run in Google Colab (or Jupyter), using widely adopted libraries such as Hugging Face Transformers, PyTorch, and vector databases, alongside hosted models (e.g., OpenAI, Gemini). Examples are deliberately scoped so that processes like fine-tuning, RAG, and agentic workflows execute quickly within session time limits, keeping the focus on practical application rather than syntax or mathematical derivations.

Software and Hardware Requirements

All session materials are delivered as a ready-to-run Google Colab notebook; participants only need a laptop with a web browser and a free Google account no local installation is required. A Jupyter version of the notebook will also be provided for anyone who prefers to run it locally. All libraries used are open-source and free, and Colab provides a free GPU runtime sufficient for the scoped fine-tuning and RAG examples. For the hosted-model portions, we will make a small number of API calls, so participants are asked to create an OpenAI account in advance and add approximately \$10 in credit. No specialized hardware is required.

FULL SCHEDULE

MONDAY, OCTOBER 5

Integrating Generative AI into R Workflows: From APIs to Shiny Apps Training Session (4 hour)

1:00 PM – 5:00 PM

Conference Center: Commonwealth 1

Presenter(s): Christopher Runyon, National Board of Medical Examiners

Abstract:

Learn to integrate large language models into R workflows through hands-on practice. Starting with architecture fundamentals, participants understand the why behind best practices before implementing various prompt engineering techniques, API interactions, multi-agentic systems, and LLM-powered Shiny applications. Moderate R fluency required; no LLM experience assumed.

Session Description: This workshop addresses the rapidly growing need for measurement professionals who only had psychometric or basic statistical training to integrate AI capabilities into R-based workflows. Educational measurement professionals are increasingly expected to incorporate AI into established workflows; however, most AI training assumes software engineering backgrounds or focuses on consumer applications, leaving a critical gap.

This workshop specifically bridges that gap by demonstrating how GenAI can augment core psychometric practices. Participants will learn to use AI in developing content, automated scoring of complex responses (e.g., applying a rubric to a performance assessment), developing novel assessment formats, and facilitating internal workflows, all while maintaining the rigor and standards necessary for prudent educational measurement. Each of these components is accompanied by a hands-on activity, where workshop attendees walk through pre-developed exercises tailored to the specific course content.

Furthermore, the emphasis on understanding architectural foundations prepares participants to make informed decisions about AI integration in high-stakes environments where incorrect implementation could affect student outcomes or institutional decisions. This foundational knowledge enables measurement professionals to become informed consumers and implementers of AI technology rather than passive users of black-box solutions.

Learning Objectives:

By the end of this workshop, participants will be able to:

- Explain key LLM architectural features that inform effective integration practices
- Apply prompt engineering principles to achieve consistent, reliable outputs in R workflows
- Implement both single-use and conversational API interactions with LLMs from R
- Design and deploy simple multi-agentic systems for complex tasks
- Build interactive Shiny applications that leverage LLM capabilities for end-user functionality

Workshop requirements:

- R (version 4.4 or higher recommended)
- R Studio
- GenAI model API key
 - o Participants will be required to set up an Anthropic API account with a \$5 credit for the workshop, which I will offer to reimburse.

Brief pre-workshop instructions on obtaining and testing their API key usage will be sent to workshop participants prior to the workshop. I will also offer to meet with participants early to ensure they are fully prepared to engage in the activities.

FULL SCHEDULE

MONDAY, OCTOBER 5

Sponsored Training Session: How Should We Benchmark AI Tutors? Toward Valid and Impact-Oriented Evaluation

Training Session (4 hour)

1:00 PM – 5:00 PM

Conference Center: King's Garden 3

Presenter(s): Rene Kizilcec, Clayton Cohn, Cornell University; Joshua Marland, Cornell University; Kirk Vanacore, Cornell University; Zhuqian Zhou, Cornell University

Abstract

As AI tutoring is expanding rapidly, so has AI tutor benchmarking, raising many questions for the measurement community: How should we benchmark tutoring quality when the target construct is latent and its most consequential outcomes are distal and often only loosely connected to the behaviors that current benchmarks reward? Existing benchmarks assess important components of tutoring, including pedagogical knowledge, error diagnosis, Socratic questioning, scaffolding, adaptive explanations, and multimodal interpretation of student work. Yet these approaches raise fundamental questions about construct definition, scoring, validity, generalizability, and the evidence needed to interpret benchmark performance as meaningful pedagogical quality. This half-day symposium and working session brings together measurement researchers, learning scientists, benchmark developers, tutoring providers, and AI developers to examine these questions. The session begins by framing AI tutor benchmarking as a measurement problem and features invited perspectives from developers of current benchmarks. Participants will critique current approaches and examine concrete design decisions, including task design, scoring rules, expert and automated evaluation, simulated students, and validation against student learning outcomes. The session aims to advance the field's discussion of valid benchmarking approaches by synthesizing priorities and establishing a community of practice around impact-oriented AI tutor evaluation.

Summary

Benchmarking has long enabled researchers and developers to compare machine-learning systems and track progress on relatively well-defined tasks with established correct answers. The rise of general-purpose pretrained models has expanded benchmarking to broad and abstract capabilities, including reasoning, honesty, helpfulness, and pedagogical quality. This shift creates a central challenge for the measurement field: performance on a particular benchmark may be treated as evidence of general competence even when the relationship between the task and the capability it supposedly measures is unclear (Raji et al., 2021).

This challenge is especially salient for AI tutoring. Tutoring quality is a latent, multidimensional, and context-dependent construct, and its most consequential outcomes—student learning, engagement, and agency—are often distal. Effective tutoring involves interpreting student thinking, diagnosing misconceptions, deciding when to explain, question, scaffold, or challenge, and adapting to students' needs over time. A response that is effective for one learner or moment may be inappropriate for another. Consequently, benchmark scores reflect not only model behavior but also designers' assumptions about what good tutoring is and what evidence should count as success (Jacobs & Wallach, 2021).

Recent benchmarks assess important components of AI tutoring, including pedagogical knowledge, error diagnosis, Socratic questioning, feedback, scaffolding, adaptive explanations, and multimodal interpretation of student work (Lane & Kageler, 2026; Lelièvre et al., 2025; Liu et al., 2025; Macina et al., 2025; Maurya et al., 2025; Otero et al., 2025; Srinivasa et al., 2025; Weitekamp et al., 2025; Yang et al., 2026). These efforts represent substantial progress but raise familiar measurement questions: What construct does each benchmark assess? Do its tasks adequately represent that construct? How should expert, student, and automated judgments be combined? How should disagreement be interpreted? Do scores generalize across students, subjects, and interaction contexts? Most importantly, what evidence is required before benchmark performance can be interpreted as evidence that an AI tutor supports student learning?

This half-day symposium and working session will situate AI tutor benchmarking within principles of educational and psychological measurement, including construct definition, validity arguments, reliability, generalizability, and consequential validity. Using concrete examples from existing and work-in-progress benchmarks, invited presenters and participants will examine how choices about tasks, rubrics, reference answers, raters, LLM-as-a-judge systems,

FULL SCHEDULE

MONDAY, OCTOBER 5

simulated students, and scoring rules shape the claims that can be made about AI tutors. Discussion will emphasize practical evaluation decisions rather than software syntax or mathematical formulas.

The session will culminate in a field-building discussion to identify shared priorities for AI tutor benchmarking. Participants will synthesize perspectives on construct definition, validity evidence, scoring, generalizability, and connections to student outcomes to articulate a preliminary agenda for valid and impact-oriented AI tutor evaluation.

Approximate balance: 40% framing/invited perspectives, 30% co-design, and 30% discussion/community building.

Text Classification with Large Language Models: Pipelines, Fine-tuning, and Measurement Validity Training Session (4 hour)

1:00 PM – 5:00 PM

Conference Center: King's Garden 1

Presenter(s): Brittney Hernandez, University of Connecticut; Kylie Anglin, University of Connecticut; Claudia Ventura, University of Connecticut

Abstract:

This interactive workshop covers methods of binary text classification using large language models: API calls to non-locally hosted models, local models, and finetuned models. Participants will build text-classification pipelines with an emphasis on practicalities and measurement validity. Tradeoffs in approaches (cost, performance, privacy, etc.) will be discussed throughout.

Session Description: Learning Objectives

- Learning Objective 1: Participants will construct a text classification pipeline to make calls to third-party LLMs (e.g., OpenAI, Claude, Gemini).
- Learning Objective 2: Participants will construct a text classification pipeline to make calls to local LLMs (e.g., Llama, Qwen).
- Learning Objective 3: Participants will identify best practices for unbiased and precise evaluation of model alignment with human labels.
- Learning Objective 4: Participants will use systematic prompt engineering to improve model performance.
- Learning Objective 5: Participants will learn a fine-tuning workflow and leave with adaptable code to improve model performance on their own data.

Content and Relevance

Unlike traditional supervised machine learning (ML), which require large, labeled datasets, extensive feature engineering and task-specific model training, LLMs offer a more accessible and high-performing alternative for text classification when labeled data are limited (Anglin et al., 2025). Pre-trained on enormous corpora, they arrive at a new task with already-encoded linguistic patterns, reducing the number of labeled examples needed for high accuracy (Devlin et al., 2019). LLM-based text classification raises measurement validity concerns beyond traditional supervised machine learning (Anglin, 2024). For example, LLM prompts can threaten construct validity. When prompting an LLM with a classification task, changes in wording can meaningfully shift the representation of the construct (Anglin, et al., 2025).

This session covers text classification with LLMs—non-local LLMs, local LLMs, and fine-tuning, each with distinct advantages/disadvantages. API-based (non-local) models from providers such as Anthropic, OpenAI, and Google offer state-of-the-art performance with minimal setup, but require sending data to a third party. Locally hosted, open-weight models keep data in-house, improve reproducibility, and decrease cost, but typically come with reductions in efficiency and in performance (i.e., reduced agreement between model classifications and expert classifications). Fine-tuning can recover some of the performance loss. However, deploying an open-weight model locally is more involved than calling an API: it requires installing and configuring a runtime such as Ollama or HuggingFace Transformers, choosing a model that fits the available hardware, and managing memory directly. The workshop walks through each of these steps directly. Emphasis will be placed on both measurement validity and practical implementation decisions, including performance metrics, sample size planning, model selection, quantization, and hardware considerations.

Prerequisite Skills and Knowledge

- Text Classification: Conceptual understanding of text classification as a method of analysis, and/or familiarity with traditional classification methods (e.g., bag-of-words, supervised classifiers)
- LLM Mechanics: Understanding of LLMs as next-token predictors, including tokenization, and a broad sense of how pre-training data shapes model behavior.
- Programming: Proficiency in one or more programming language(s) such as Python or R. Conceptual understanding of file input/output, data manipulation, functions, and loops.
- For more info and prerequisite software/packages see: <https://github.com/hernandezb3/llm-text-classification> or email brittney.hernandez@uconn.edu.

Writing an AI-Native Dissertation

Training Session (4 hour)

1:00 PM – 5:00 PM

Conference Center: King's Garden 2

Presenter(s): *Damian Betebenner, Center for Assessment*

Abstract:

A hands-on, half-day session in which participants build their own AI-native dissertation: a reproducible project that produces a website, a thesis PDF, an interactive application, and an API queryable by AI agents. Participants leave with a personalized environment generated from pre-session research scoping responses.

Session Description: By the end of the session, participants will be able to:

- Explain the dissertation as a multi-modal knowledge-transmission platform, and articulate how an AI-native workflow changes its reach, reproducibility, and the locus of human responsibility.
- Run and navigate their own dissertation environment — an R package, Quarto sources, and a NextJS application — generated in advance from their pre-session scoping responses.
- Render a single source to three outputs: an interactive website, a formatted thesis PDF, and a working paper.
- Apply the write-render-review cycle to draft and revise dissertation content with AI as a collaborator, while retaining full editorial control.
- Link narrative prose to live analysis so that reported values regenerate from code rather than being transcribed by hand.
- Build and publish their dissertation to the open web: the Quarto site as a public website and the interactive application deployed to Vercel.
- Expose their research through a REST API and Model Context Protocol (MCP) tools so that AI agents can query their analyses and data directly.
- Develop practical fluency in advanced AI tool use — a skill distinct from understanding how the models work internally — including customizing the AI harness (project rules, context files, and MCP tools) to a particular scholarly purpose.
- Document AI use transparently and apply emerging norms for attribution, reproducibility, and institutional policy.

Content and Relevance

The dissertation is, at root, a technology for transmitting understanding from one mind to others — and for six centuries its fundamental mode has barely changed. Knowledge passed from the oral disputation of the medieval university, to hand-copied manuscripts, to the printing press, to the PDF; each step widened reach but preserved the same modality: one human writes, and one human reads, in linear text. The digital era digitized the container, not the knowledge — a PDF is a photocopy of a printed page, with data and code still hidden “available upon request.” Measured against how scholarship is now produced and consumed, that mode is increasingly outdated. Two developments change the equation: APIs make findings addressable rather than merely readable, and AI agents mean the consumers of research are no longer exclusively human.

This session teaches a concrete, open, forkable framework that treats the dissertation as a platform: content and analysis are authored once and, driven by a single declarative specification and shared content layer, surfaced as a website, a thesis PDF, a working paper, a REST API, MCP tools, and an interactive application. The architecture is

deliberately ordinary underneath (a standard R package, Quarto, a NextJS app, Git), so participants learn transferable practices rather than a bespoke tool. The environment is ready-to-hand — the tooling recedes so the author dwells in the dissertation, not the interface. The recurring discipline is “leave the technical plumbing to the AI, and own the content”: research questions, analysis, and interpretation remain the scholar’s responsibility. The context an author invests up front — framing, advisor writings, key literature, style exemplars — becomes an AI-queryable knowledge layer, because AI’s usefulness compounds with the depth of context it can draw on.

Fluency with AI is fast becoming an essential professional skill for measurement professionals, and there is no better place to build it than an extended, high-stakes scholarly project such as a dissertation or thesis, where genuine judgment — not toy exercises — is on the line. A central aim of the session is therefore practical, advanced tool use — directing AI well and customizing its harness of project rules, context files, and MCP tools to scholarly ends — a competence distinct from understanding how the models work internally.

The session operationalizes the measurement-science values at the center of the conference theme. Reproducibility underwrites the validity of every reported number, because results regenerate from versioned code with fixed seeds. Transparency follows from open, auditable analysis functions and logged AI interactions. Whatever the system derives or generates on the author’s behalf — the project specification, the scaffolding, AI-suggested prose — is inspectable, contestable, and exportable, so human agency is preserved and AI remains a collaborator and an access layer, never an author. Attribution and disclosure are built into the workflow. The framework and its onboarding are open source; the session is about the practice of AI-native scholarship, not any product.

Software needed

- A modern web browser
- Git
- R (≥ 4.5)
- Quarto (≥ 1.6)
- Node.js (LTS) with pnpm
- a free GitHub account
- optionally, a free Vercel account for deployment.
- RStudio or VS Code is recommended as an editor.
- Most importantly, an AI coding assistant (for example, Claude Code, Cursor, or Codex) — the tool participants point at their repository to make changes throughout the session.
- With the single exception of the AI assistant, all required software is free and open-source.

Installation / payment

Participants install the above ahead of time using setup instructions emailed before the session, and complete the pre-session scoping questionnaire. Every component except one is free: Git, R, Quarto, Node.js/pnpm, and the GitHub and optional Vercel accounts cost nothing. The one paid requirement is the AI coding assistant — the engine that reads and updates the dissertation repository. Sustained, agentic use during the session realistically requires either an active paid subscription (for example, Claude Pro/Max, ChatGPT Plus, or Cursor Pro— roughly \$20/month) or pay-as-you-go API credits from a provider such as OpenAI or Anthropic. Free tiers exist but are typically too rate-limited for hands-on work, so participants should arrive with an active paid plan or a small amount of API credit already configured.

Hardware

A laptop (macOS, Windows, or Linux) on which the participant can install software (administrator rights), with roughly 5 GB of free disk space and reliable Wi-Fi.

No specialized hardware is required.

FULL SCHEDULE

TUESDAY, OCTOBER 6

What Are We Assessing in Human-AI Interaction? Keynote: Dragan Gašević

Keynote

8:00 AM – 9:00 AM

Conference Center: Grand Ballroom 1

Session Description: Artificial intelligence is changing what observed performance can tell us about human learning. People can produce better outcomes with AI without developing the knowledge and skills needed to perform beyond its assistance. This growing separation between performance and learning creates a fundamental problem for assessment. If assessment continues to privilege successful task completion, it risks measuring what people can accomplish with AI rather than the human capability developed through that interaction. This talk argues that assessment in human-AI interaction needs to provide evidence of durable human capability while recognizing that productive use of AI is itself becoming part of competent performance. Metacognition is central to this challenge because learners need to regulate what they delegate to AI and what they need to develop themselves. Drawing on research in human-AI interaction, learning analytics, and assessment, the talk considers how assessment can support more valid inferences about human learning and capability from performance that is increasingly produced through human-AI interaction.

FULL SCHEDULE

TUESDAY, OCTOBER 6

Keynote Overflow Room

Keynote

8:00 AM – 9:00 AM

Conference Center: King's Garden 5

Break

9:00 AM – 9:15 AM

AI for Alignment to Standards

Individual Paper Session

9:15 AM – 10:45 AM

Conference Center: Brigade

Presentations:

- **Evaluating Standards Alignment and Uniqueness of AI-Generated Assessment Items**

Presenter(s): Vandhana Mehta, Pearson; Laine Bradshaw, Pearson

This study proposes an evaluation framework for AI-assisted assessment item development. Combining expert reviews evaluated via Gwet's AC2 with NLP similarity metrics, we compare prompt-from-scratch and reverse prompt engineering methodologies. We establish pilot-derived validation guardrails to optimize item diversity, prevent cloning, and ensure precise diagnostic attribute alignment.

- **AI-Enabled Quality Assurance for Multiple-Choice Assessment Items**

Presenter(s): Steven Moore, George Mason University; Nick Diana, Colgate University

Generative AI can scale multiple-choice question production without comparable advances in quality assurance. This work-in-progress review synthesizes recent AI-enabled methods for evaluating MCQ quality, anchored on the 19-criteria Item-Writing Flaws rubric. It traces the field from flaw detection through feedback and revision to psychometric validation and NLP benchmark auditing.

- **Predicting Item-to-RPLD Matches with Structured AI Reasoning**

Presenter(s): Kang Xue, Cambium Assessment; Christina Schneider, Cambium Assessment, Inc.; Sangdon Lim, Cambium Assessment; Garron Gianopulos, Cambium Assessment; Alexandra Perez, Cambium Assessment, Inc.

Large language models are increasingly used to support assessment design, yet their ability to classify items into Range Performance Level Descriptors remains unclear. This study compares small and large language models for item-to-RPLD matching and examines structured reasoning, ensemble classification, and AI-human agreement within a Principled Assessment Design framework.

- **Can AI Detect Item-Metadata (Mis)alignment? An Empirical Study Using State Assessment Data**

Presenter(s): Hye-Jeong Choi, HumRRO; Meng Fan, HUMRRO; Yufeng Berry, Minnesota Department of Education; Emily Borawski, HumRRO; Ruixue Liu, Fairfax County Public Schools, VA

This study empirically examines the extent to which large language models (LLMs) can classify assessment items according to standards specified in official blueprints. Using released Smarter Balanced items, agreement between LLM and blueprint classification will be evaluated across three LLMs to assess their potential for assessment quality assurance.

- **Achievement-Level Descriptor Conditioned Option-Selection Probability Modeling with Large Language Models**

Presenter(s): Yuxiao Zhang, Purdue University; Sandra Sweeney, Cognia; Rachael Ahn, Cognia; Louis Roussos, Cognia

Using released MCAS Grade 10 ELA items, this study proposes achievement-level descriptor-conditioned option-selection probability modeling for LLM-supported item evaluation and compares it with direct response generation. The proposed approach outperformed the baseline in score-level and item-difficulty recovery, suggesting a more stable formulation for AI-assisted item evaluation.

FULL SCHEDULE

TUESDAY, OCTOBER 6

Comparing and Exploring Human and AI Scoring
Individual Paper Session
9:15 AM – 10:45 AM
Conference Center: Ft. Pitt

Presentations:

- **Evaluating Machine Learning, Transformer, and LLM Approaches for Large-Scale Automated Essay Scoring**
Presenter(s): Cyrus Sai-Cheong Chan, Vretta; Charles Anifowose, Vretta; Su Zhang; Winnie Wing-Yee Tse, Education Quality and Accountability Office; Nizam Radwan, Education Quality and Accountability Office; Hyunah Kim, Education Quality and Accountability Office
This study compared machine learning, transformer fine-tuning, and prompt-based LLM approaches for automated essay scoring in a Canadian large-scale provincial assessment. Machine learning and fine-tuned transformers achieved the highest reliability, while prompt-based LLMs provided greater interpretability. Ensemble architectures improved LLM performance, highlighting opportunities for balancing accuracy and explainability.
- **Two Uses of Human-scored Anchors: Few-shot Direct Scoring and Anchor-based Comparative Grading**
Presenter(s): Prashreet Poudel, Western Carolina University; Huayue Gu, Kennesaw State University; Collin Lynch, North Carolina State University; Zhikai Gao, Western Carolina University
This study evaluates direct few-shot scoring and anchor based comparative judgment for LLM essay grading. Using human scored anchor essays, we compare their accuracy and stability. Although both approaches produce competitive scores, comparative judgments demonstrate better stability and consistency, suggesting it as a better alternative for grading writing assessment.
- **Hidden Judgment: Detecting Latent Scoring Decisions in LLM Essay Rating Rationales**
Presenter(s): Joe Grochowalski, College Board; Claudia Ventura, College Board
LLM essay raters act on scoring judgments their explanations might not reveal. We embed the rationales Sonnet 4.6 produced for 500 essays and inspect scoring decisions where explanation of judgments are near-identical. This surfaced distinctions the rationale failed to encode, which is a reusable diagnostic for hidden judgments.
- **Minimal-Pair Perturbation of LLM Essay Scorers: What Triggers Invalid Classifications?**
Presenter(s): Ryan Lerch, UCLA
Automated essay scorers sometimes reject responses as invalid. What essay features cause these decisions? Extending our NCME 2025 invalidity-classification work, we use minimal-pair perturbations to isolate causal triggers across three LLMs and a fine-tuned encoder on essays from the ASAP dataset. This minimal-pair framework also extends to individual score points.
- **Seven Ways to Cut a Score: Mapping Essay-Scoring Predictions to Ordinal Ratings**
Presenter(s): Ahmed Bediwy, Pearson; Martha Bellows, Pearson; Susan Lottridge, Pearson
Automated essay scoring often predicts continuous scores that must be converted to ordinal rubric levels. We compare seven cutpoint estimation methods on ASAP 2.0 using a Longformer model. Results reveal an accuracy–distributional fidelity trade-off: QWK optimization maximizes agreement, while balance optimization best reproduces human score distributions.

Connecting LLMs to Digital Assessment & Scoring**Individual Paper Session****9:15 AM – 10:45 AM****Conference Center: Heinz****Presentations:**

- **AI-Human Writing Score Agreement in Primary Grades: A Longitudinal Investigation**

Presenter(s): Young-Suk Kim, University of California Irvine; Tamara Tate, UC Irvine; Ali Keramati, University of California Irvine; Julie Jung, Harvard University; Max Lu, Harvard Graduate School of Education; Justin Cheok

This longitudinal study examines AI-human writing score agreement for students in kindergarten through grade 4 across narrative, opinion, and expository genres. Two AI tools (GPT-5.5 and a multi-agent debate system) are compared to human scores on quality and CBM measures, examining variation by grade, genre, and prompting approach.

- **Composite Scores vs. Preference Rankings: Measuring Architecture Effects in LLM Feedback**

Presenter(s): Harvey Ngoe Kolle, University of Alberta; Carrie Demmans Epp, University of Alberta; Amna Liaqat, George Mason University; Maria Cutumisu, McGill University

AI-generated feedback is often judged by quality ratings or preference rankings but rarely both. Comparing a multi-agent system, a single-agent system, and human feedback on student writing, we show that the two measures can disagree, which has consequences for how AI-generated feedback should be evaluated.

- **Task Formulations in LLM Prompts: Implications for AI-Driven Automated Assessment**

Presenter(s): Amy Adair, ETS; Field Watts, ETS; Lei Liu, ETS; Libby Gerard, University of California, Berkeley; Marcia Linn, University of California, Berkeley

This study investigates how task formulation (i.e., ways of defining the idea detection task within a prompt to a large language model) affects automated scoring of middle school students' science writing. Three formulations are compared, with implications for the design of more reliable and scalable AI-driven formative assessment systems.

- **Keystroke Dynamics in Writing: Subgroup Analysis of Processes and Scoring Features**

Presenter(s): Chen Li, ETS; Mo Zhang, Educational Testing Service; Xiang Liu, ETS

We analyze keystroke dynamics from a large-scale writing assessment using transformer models to characterize writing processes and compare subgroup differences across demographics, score levels, and testing modes. Findings identify performance-related patterns, enable targeted feedback, and provide process-based evidence to support validity, fairness, and interpretable AI in writing assessment research applications.

- **Connecting Student Feedback Understanding to Gains in Argumentative Writing and Revision**

Presenter(s): Diane Litman; Zhexiong Liu, University of Pittsburgh; Tianwen Li, University of Pittsburgh; Lindsay Clare Matsumura, University of Pittsburgh; Richard Correnti, University of Pittsburgh; Elaine Wang, Rand Corporation

While automated writing evaluation (AWE) systems typically assess essay quality, we develop a process-oriented approach assessing revision alignment with feedback. Classroom deployments show that student perceptions of AWE feedback correlate with writing improvement. Furthermore, feedback designed to build revision knowledge can enhance student outcomes, shifting focus from product to process.

FULL SCHEDULE

TUESDAY, OCTOBER 6

Construct Representation and Validity Evidence in Automated Scoring Systems

Coordinated Paper Session

9:15 AM – 10:45 AM

Conference Center: Smithfield/Birmingham

Session Description:

This coordinated session brings together five studies that examine automated scoring systems through the lens of measurement theory and validity evidence. As AI-based scoring systems are increasingly used to evaluate complex student responses, traditional evidence based on prediction accuracy or agreement with human scores is no longer sufficient. Stronger validity arguments require attention to how scores are produced, what constructs are represented, how human ratings function as criteria, whether scoring models rely on construct-relevant evidence, and whether AI systems reproduce or introduce measurement-related bias. The first paper presents a scoping review of how human ratings are used in AI-based automated scoring systems and considers implications for validity arguments. The second examines representation-level evidence in a BERT-based automated essay scoring model to investigate what information is encoded in scoring-relevant model representations. The third uses controlled rewriting to evaluate whether automated scoring systems rely on construct-relevant language features or linguistic shortcuts that may threaten fairness. The fourth examines rater effects in a large language model-based short-answer scoring, focusing on severity and centrality as measurement concerns. The fifth extends the discussion to automated generation and scoring of Maze reading comprehension assessments.

Discussant: Jinnie Shin (University of Florida)

Chair: Jinnie Shin (University of Florida)

Presentations:

- **Human Ratings in AI-Based Automatic Scoring Systems: A Scoping Review**
Presenter(s): Yue Xu, University of Florida; Anna Pauline Aguinalde, University of Florida; Xiaomeng Xiong, University of Florida; Corinne Huggins-Manley; Jinnie Shin, University of Florida
- **Explaining Automated Essay Scoring: A Representation-Level Analysis of a BERT-Based Model**
Presenter(s): Lingchen Kong, University of Florida; Jinnie Shin, University of Florida
- **When Language Becomes a Shortcut: Assessing Construct Validity and Fairness in Scoring**
Presenter(s): Xiaomeng Xiong, University of Florida; Corinne Huggins-Manley; Jinnie Shin, University of Florida; Christan Grant, University of Florida
- **Disentangling Severity and Centrality: Rater Effects in LLM-Based Automated Short-Answer Scoring**
Presenter(s): Xiaomeng Xiong, University of Florida; Corinne Huggins-Manley; Jinnie Shin, University of Florida
- **Automated Generation and Scoring of Maze Reading Comprehension Assessments**
Presenter(s): Hatice Kubra Karakis, University of Florida; Walter Leite, University of Florida; Logan Scott, University of Florida; Xinyi Tai, University of Florida; Akihito Kamata, Southern Methodist University

FULL SCHEDULE

TUESDAY, OCTOBER 6

Evaluating AI-Generated NAEP items: The Assessment Developer's Perspective
Coordinated Paper Session
9:15 AM – 10:45 AM
Conference Center: King's Garden 5

Session Description:

This coordinated session presents focused evaluations of Wilbur, an AI-enabled item generation tool, for National Assessment of Educational Progress (NAEP) Reading, Mathematics, and Science, framed by a broader validity argument about what counts as evidence for AI-generated assessment items. Across studies, assessment developers working on NAEP generated, reviewed, and evaluated AI-produced draft items under structured, intentionally constrained conditions, focusing on whether items met NAEP expectations for quality, alignment, fairness, and accuracy.

Findings show that AI-generated items were often usable as early drafts, but limitations varied by subject. In Reading, quality was heavily dependent on passage and construct demands. In Mathematics, newer model versions produced stronger drafts, though expert revision remained necessary. In Science, Wilbur performed comparably to outside item writers (OIWs) and other AI tools in constrained comparisons, but items were less successful in fully integrating science practices, content, and crosscutting dimensions.

Across all three areas, assessment developers determined that no items were ready for operational use without human revision, and the findings reflect expert-judged draft quality only, without student-response data. Together, the session shows that AI can support expert-led early-stage item development while underscoring the continued vital role of assessment developers in ensuring validity, quality, and responsible use.

Discussant: Terran Brown
Chair: Cara DiMeglio (ETS)

Presentations:

- **From Expert Approval to Validity Evidence: Evaluating AI-Generated Assessment Items**
Presenter(s): Chris Meador, ETS
- **Evaluating an AI-Item Generation Tool for NAEP Reading**
Presenter(s): Cara DiMeglio, ETS
- **Evaluating an AI-Item Generation Tool for NAEP Mathematics**
Presenter(s): Cheryl Van Ness, ETS; Karen Parker, ETS; Illona Minchuk, ETS
- **Evaluating an AI-Item Generation Tool for NAEP Science**
Presenter(s): Carolina Safar, ETS; Mitchell Price, ETS; Jeffery Ackley, ETS

FULL SCHEDULE

TUESDAY, OCTOBER 6

Industry Leader Perspectives: How AI is Changing K12 Measurement Roles

Panel Session

9:15 AM – 10:45 AM

Conference Center: Commonwealth 2

Session Description:

The assessment industry is undergoing a significant transformation driven by the rapid integration of artificial intelligence into both daily workflows and core measurement practices. Professionals are increasingly leveraging AI to enhance tasks such as coding, technical writing, and research, while organizations are embedding AI across the assessment lifecycle—from item development and scoring to calibration and standard setting. This shift is further accelerated by organizational leadership priorities to improve efficiency, reduce costs, and modernize assessment systems. Additionally, job expectations are being modified to explicitly require AI-related competencies for both current professionals and new entrants.

Despite widespread adoption, the pace and nature of this change vary across organizations, and our understanding of the foundational competencies for measurement professionals in an AI-enabled landscape are still emerging. This panel brings together senior leaders from diverse K-12 assessment organizations spanning summative, interim, formative screener and college readiness to share perspectives on how roles, skills, and research are evolving. Panelists will discuss practical impacts on day-to-day work, changing job requirements, and how AI innovations in measurement are changing the work of psychometricians. The session emphasizes interactive dialogue, combining guided discussion with audience participation to explore the future of the profession.

Moderator: Moderator: Susan Lottridge (Pearson)

- **Industry Leader Perspectives: How AI is Changing K12 Measurement Roles**

Presenter(s): Catherine Close; Kristen Huff, Curriculum Associates; Dianne Henderson; Jennifer Dunn, The College Board; Christina Schneider, Cambium Assessment, Inc.

Sponsor Showcase

9:15 AM – 10:45 AM

Conference Center: Commonwealth 1

Synthetic Students, Responses & Simulation

Individual Paper Session

9:15 AM – 10:45 AM

Conference Center: Duquense

Presentations:

- **It Quacks Like a Response: Evaluating Synthetically-Generated Responses in Assessment Research**

Presenter(s): Cole Walsh, Acuity Insights; Rodica Ivan, Acuity Insights

This study examines the quality of synthetic open-response data generation for assessment development and psychometric analyses and suggests methods for measuring synthetic response fidelity and diversity. We treat existing real response data as ground truth and evaluate the effectiveness of conditioning on real test taker responses when generating synthetic data.

- **Do Synthetic Students Thrive? Validity and Fairness of Generated Assessment Data**

Presenter(s): Neba Nfonsang, Oluko Learning; Temple Lovelace, Assessment for Good, AERDF

Synthetic assessment data can match real distributions while losing the within-group spread that asset-based instruments reveal in learners furthest from opportunity. Using a stratified, majority Black and Latinx sample, we test whether our generation pipeline preserves construct structure or pulls learners toward their group averages, and report an LLM-to-KNN finding.

- **Rethinking Pilot Data: Evaluating LLM Synthetic Data for Scale Development**

Presenter(s): Margarita Olivera Aguilar, American Institutes for Research; Samuel Rikoon, American Institutes for Research; Paul Bailey, American Institutes For Research; Michael Kruse, American Institutes for Research; Kamal Middlebrook, American Institutes for Research

We evaluate whether LLM-generated synthetic data can support early psychometric analysis of K-12 measures. Testing 29 prompts across survey datasets, results show limited success generating usable data, but also (when successful) promising recovery of model fit and structural relationship. Item-level accuracy remains weak, suggesting cautious, targeted applications.

- **Simulating Examinee Responses to Figure Matrices Items Using VLMs for Item Evaluation**

Presenter(s): Onur Demirkaya, Riverside Insights; Evelyn Johnson, Riverside Insights

This study evaluated whether visual embeddings from SigLIP and DINOv2 could be used to simulate responses to Figure Matrices items. Models predicted response probabilities using item screenshots and examinee ability estimates. Results showed moderate probability recovery but limited item difficulty recovery, indicating promise for early screening but not calibration replacement.

Using LLMs to Assist in IRT Modeling

Individual Paper Session

9:15 AM – 10:45 AM

Conference Center: King's Garden 4

Presentations:

- **Prior-Informed 3PL Calibration: Reducing Sample Size via Predictive Modeling**

Presenter(s): Ruitao Liu, ACT ; Aixin Tan, University of Iowa

This simulation study evaluates whether predictive item-parameter priors can reduce respondent requirements for 3PL IRT calibration. Across 27 prior-quality conditions and seven sample sizes (400--1600), calibration at 400 respondents matched a 2000-respondent baseline only when guessing and discrimination prior errors were each reduced by at least 5--10%, yielding 80% sample-size

- **Posterior-Based Item Fit Diagnostics for Variational IRT Models**

Presenter(s): Yong Luo, Pearson; Yawei Shen

Item fit diagnostics are essential for operational IRT but have no established analogue for neural IRT models. We develop a per-item fit diagnostic from ELBO reconstruction error, validate it against $S-X^2$ across sample sizes and test lengths, and introduce a person-fit diagnostic from the variational posterior KL divergence.

- **The Beta-Bernstein Bridge: Closed-Form Bayesian Measurement Beyond Monotone Item Response Functions**

Presenter(s): Yumou Wei, Carnegie Mellon University

Item response theory forces a choice between flexible curves and tractable inference. We bridge them: writing any bounded item response function as a Bernstein polynomial makes the Beta prior conjugate, giving an exact closed-form ability posterior as a Beta mixture. It reproduces IRT and recovers non-monotone curves with bimodal posteriors.

- **Constrained Symbolic Regression for Interpretable IRT Diagnostics**

Presenter(s): Sirui Wu, University of British Columbia; Guanyu Chen, University of British Columbia

We adapt symbolic regression ideas to a constrained post-baseline diagnostic workflow for dichotomous logistic IRT models. CSME searches prespecified logit-scale residual terms, selects by validation loss with a complexity penalty, and evaluates held-out performance. Across simulations and TIMSS data, it preserves correct baselines and produces interpretable follow-up hypotheses about misfit. Bayesian Consensus

- **Calibration of Continuously Evolving IRT Item Banks**

Presenter(s): Paul Jewsbury, Duolingo; Steven Nydick, Duolingo; Manqian Liao, Duolingo; Siyuan Chen, Duolingo, Inc.

AI-generated items with NLP-predicted parameters produce ever-larger, continuously-evolving banks, outpacing MCMC estimation of explanatory IRT models. We propose Bayesian consensus calibration, a divide-and-conquer method that fits each calibration period on disjoint data, links posterior draws while propagating uncertainty, and aggregates them into a consensus posterior with a single consensus prior.

Automated Item Generation & Item Writing

Individual Paper Session

11:00 AM – 12:30 PM

Conference Center: Heinz

Presentations:

- **Efficacy of Using Student-AI Co-Authored Math Word Problems in an ITS**

Presenter(s): Kole Norberg, Carnegie Learning; April Murphy, Carnegie Learning; Stephen Fancsali, Carnegie Learning, Inc.; Steven Ritter, Carnegie Learning

We evaluated the generalizability of math word problems co-authored by students with the help of generative AI. 1,361 students solved and rated co-authored and standard problems in an Intelligent Tutoring System. Performance on co-authored problems was comparable to standard problems, modestly reducing error rates. Ratings revealed diverse scenario preferences.

- **Data-driven AIG process: Lessons from CTT-IRT Discrepancies**

Presenter(s): NaRa Nam; Euigyum Kim, Sogang University; Hyo Jeong Shin, Sogang University

This study examines a data-driven AIG process by leveraging CTT-IRT difficulty discrepancies as psychometric indicators for validating AIG frameworks. Using assessment response data, we identified three empirically derived item features associated with classification mismatches. The results suggest that such discrepancies can reflect structural properties relevant for improving AI-generated item validity.

- **Generative Drafts as a Tool for Expert Item Writers**

Presenter(s): William Muntean, National Council of State Boards of Nursing; Joe Betts, National Council of State Boards of Nursing; Zhuoran Wang, National Council of State Boards of Nursing; Pauline Aguinalde, University of Florida

This study evaluates generative language models as a source of initial framing for underpopulated test blueprint targets. Model-framed and SME panel-framed drafts were compared after undergoing identical item writing workflows. Evidence spanning blinded source classification, draft-to-final linguistic analysis, and psychometric calibration informs expert-governed item development.

- **Responsible use of generative AI when creating reading comprehension questions: Inference matters**

Presenter(s): Zuowei Wang, Educational Testing Service; Michael Flor, ETS; Jiyun Zu, ETS; Tenaha O'Reilly; Wanjing Ma, Stanford University

AI-generated and expert-created reading comprehension questions can show similar item statistics yet differ in the types of inferences required. This difference stemmed from AI's failure to follow prompts during an intermediate item generation step. Evaluations of AI-generated items should document prompts and generation steps to identify and mitigate construct-relevant differences.

FULL SCHEDULE

TUESDAY, OCTOBER 6

Constructed-Response Scoring Evidence from LLMs, Embeddings, and Rubrics
Coordinated Paper Session
11:00 AM – 12:30 PM
Conference Center: Grand Ballroom 1

Session Description:

This coordinated paper session examines how large language models, embeddings, and rubric-aligned evidence can improve the scoring and validation of constructed responses across writing, speech, and critical-thinking tasks. The papers address complementary stages of the automated-scoring lifecycle: generating verified silver labels for fine-grained formative writing feedback; using LLM agents to refine rubrics by analyzing human–engine discrepancies; applying clustered text embeddings to identify inconsistent raters when conventional common-item designs are limited; reframing early-literacy speech scoring as contrastive comparison between student utterances and reference exemplars; and evaluating LLM-supported scoring of critical-thinking subskills in argumentative essays. Across these studies, AI is not treated simply as a replacement for human scoring, but as a tool for making scoring evidence more explicit, scalable, and diagnostically useful. Together, the papers highlight both promise and boundary conditions: strong performance for well-defined constructs and item types, persistent challenges for nuanced reasoning and subjective style judgments, and the need for psychometric validation, human adjudication, and transparent audit trails before AI-supported scoring systems are used operationally.

Discussant: Ikkyu Choi (ETS)

Chair: Edward Wolfe (University of Iowa)

Presentations:

- **Toward LLM-supported Automated Assessment of Critical Thinking Subskills**
Presenter(s): Andrew Lan, University of Massachusetts, Amherst
- **Multi-Agent LLM Annotation and Scoring for Training Fine-Grained K-12 Writing-Feedback Models**
Presenter(s): Justin Barber, Pearson; Michael Hemenway, Pearson; Martha Bellows, Pearson; Susan Lottridge, Pearson
- **LLM-Based Rubric Refinement in Multi-Agent Automated Scoring**
Presenter(s): Alexander Kwako, Cambium Assessment; Cristina Everett, Cambium Assessment
- **Contrastive Embedding Approaches for Automated Speech Scoring of Early Literacy**
Presenter(s): Gitit Kehat, Cambium Assessment; Christopher Ormerod, College Board; Zachary Schultz, Cambium Learning Group, Inc.; Debbie Dugdale, Cambium Assessment, Inc.; Alexander Kwako, Cambium Assessment
- **Putting it Together: Using Embeddings Clusters to Identify Inconsistent Raters**
Presenter(s): Trevor Breakspear, Cambridge Assessment; Abhirup Chakravarty, Cambridge Assessment; Mark Brenchley, Cambridge Assessment; Yan Huang, Cambridge Assessment

FULL SCHEDULE

TUESDAY, OCTOBER 6

LLM Judges & Human-AI Agreement in Scoring and Test Assembly
Individual Paper Session
11:00 AM – 12:30 PM
Conference Center: Ft. Pitt

Presentations:

- **A Feasibility Study on Retrieval Practice Using Large Language Models**
Presenter(s): Marcus Leong, Wesleyan University; Jennifer Rose, Wesleyan University; Lisa Dierker, Wesleyan University; Antonio Laverghetta Jr., Wesleyan University
Retrieval practice supports learning but requires educators to build large item banks. We compared LLM-generated and human-written retrieval practice items in an introductory psychology course to test whether LLM items match instructor quality. LLM items exhibited overall weaker psychometric properties, suggesting that human supervision remains necessary during item generation.
- **Auditing Human, Fine-Tuned DeBERTa, and Zero-shot GPT Essay Raters**
Presenter(s): Wei Huang, University of Alabama, College of Education; Stefanie Wind, The University of Alabama
This work-in-progress compares human, GPT, and fine-tuned DeBERTa raters for English-learner writing. We apply many-facet Rasch measurement (MFRM), differential rater functioning (DRF) to examine rating quality including rater-by-criterion interactions. We aim to investigate how different types of AI raters inherit or diverge from human rater effects.
- **Assessing Large Language Model Performance in Post-Certification-Examination Comment Categorization**
Presenter(s): Huaping Sun, American Board of Psychiatry and Neurology; Kristin O'Brien; Colleen Kave; David Shin, American Board of Psychiatry and Neurology; Qiao Lin; Jeffrey Lyness
This study evaluates a large language model (LLM) for automating post-certification-exam comment analysis. The LLM showed almost perfect agreement with human raters in sentiment analysis and substantial agreement in thematic categorization, with lower performance in subjective domains. The results highlight feasibility while underscoring limitations in contextual interpretation and refinement needs.
- **Beyond Agreement: Calibrating Automated Scoring to Human Judgment Using Control Scripts**
Presenter(s): Mark Dulhunty, Australian Council for Educational Research; Alex Codoreanu, Australian Council for Educational Research; Nathan Zoanetti, ACER
An ensemble automated scoring approach was evaluated against 260,000 responses spanning 55 constructed-response reading items. Performance was then compared to 27 human markers using 477 control scripts. The ensemble achieved perfect agreement on control scripts for 39 items (71%), matching or exceeding human marker performance and supporting operational quality assurance.

Modeling Item Cognitive Demand & Difficulty

Individual Paper Session

11:00 AM – 12:30 PM

Conference Center: King's Garden 4

Presentations:

- **Exploring AI-driven Methods for Pre-calibrating Difficulty of Mathematics Items with Visual Content**
Presenter(s): Leah Walker; Jinnie Choi, Savvas Learning Company
Use of AI is bringing efficiencies in replacing previously laborious field-testing with scaled item parameter pre-calibration, especially in language assessment. However, gaps in research exist in replicating this accomplishment for mathematics assessment items with visual content. This study explores the methodological options of using AI in pre-calibrating multimodal mathematics items.
- **Estimating Uncertainty in Text Classifier Validation**
Presenter(s): Kylie Anglin, University of Connecticut
Text classification allows researchers to extract measures from natural language data but reported performance metrics are nothing more than point estimates accompanied by uncertainty. We address confidence interval estimation under conditions common in the validation of text classifiers in education, including small validation sample sizes, rare constructs, and nested data.
- **What LLMs Capture and Miss About Cognitive Sources of L2 Item Difficulty**
Presenter(s): Meng Lyu; Moses Omojekunola, HSE University; Feng Wang, UMass, Amherst
We link item difficulty from about 1,200 examinees per item to a cognitive-diagnostic Q-matrix, asking where large language models' zero-shot difficulty estimates align with real difficulty and whether misalignment tracks the attributes models fail to recover. Pilots show strong recovery for rule-based attributes, weak alignment on meaning-dependent items.
- **Math Item Difficulty Prediction with Multimodal Input**
Presenter(s): Kai North, Cambium Learning Group, Inc.
Item difficulty prediction depends on effectively leveraging item characteristics. Many math items contain visual stimuli that are difficult to incorporate into prediction models. We evaluate open-source multimodal large language models on predicting the difficulty of math items with visual content and find they can successfully estimate item difficulty.
- **Beyond Item Text: Modeling Latent Cognitive Demand for Math Item Parameter Modeling**
Presenter(s): Jae Jun Jong; Suhwa Han, Cambium Assessment Inc.
This study investigates whether modeling latent cognitive demand can improve math item parameter prediction beyond item content alone. Leveraging Schoenfeld's episode theory and standards-based learning progressions, we develop and evaluate three encoder modules to quantify the contribution of each component.

Novel Approaches to Scoring in Mathematics and Computer Science**Individual Paper Session****11:00 AM – 12:30 PM****Conference Center: Smithfield/Birmingham****Presentations:**

- **Machine-Verifiable Item Models for Geometry Diagram Generation**

Presenter(s): Mei Chen, Renaissance Philanthropy; Gordon Stein, Renaissance Philanthropy

Language models can draw diagrams that look right but are geometrically wrong. We have them specify constructions, not coordinates; a symbolic engine computes exact figures and checks geometric predicates, enabling automatic item generation with computed answer keys. Structured generation improves pass rates; human validation reveals limits in checker completeness.

- **Fine-Tuning Large Language Models for Codebook-Guided Coding of Student Metaphors About Mathematics**

Presenter(s): Liang Zhang, University of Michigan; Jinfa Cai; Stephen Hwang, University of Delaware; Yue Ma

Using 2,265 human-coded Grade 6–8 mathematics metaphor responses, this study compares proprietary prompt-only LLMs with open-weight models before and after LoRA fine-tuning. Fine-tuning substantially improved coding accuracy and reliability, often surpassing proprietary models, supporting scalable, locally controlled, privacy-conscious measurement of students' mathematical attitudes, beliefs, and experiences from their metaphor responses.

- **Can LLMs Replace Semantic Similarity for Scoring Student Code Explanations?**

Presenter(s): Arun Balajee Lekshmi Narayanan, University of Pittsburgh; Mohammad Hassany; Peter Brusilovsky, University of Pittsburgh

We compared LLM-based scoring with semantic similarity for assessing student self-explanations of code, framed as binary classification. Using generative AI to synthesize negative examples and balance a skewed dataset, we find LLM prompting ($F1 = 0.98$) significantly outperforms semantic similarity ($F1 = 0.72$), with fewer false negatives.

- **Speed–Accuracy Trade-offs in Knowledge Component Identification from Student Code**

Presenter(s): JING FAN, Aalto University; Tsvetomila Mihaylova, Aalto University; Peter Brusilovsky, University of Pittsburgh; Juho Leinonen, Aalto University; Charles Koutchme, Aalto University; Narges Norouzi, University of California, Berkeley; Bitu Akram, North Carolina State University; Arto Hellas, Aalto University

This study compares feature-based machine learning and Gemma 3 models for identifying knowledge components in student code. Traditional lexical models achieve higher accuracy and substantially lower inference time, while larger language models perform better than smaller ones. Findings inform scalable AI-supported assessment and feedback in programming education.

- **LLM-As-An-Assessor: Can Open-Weight LLMs Assess Computational Thinking via Student-Designed Embodied Games?**

Presenter(s): Will Lee, UMass Amherst; Ivon Arroyo, UMass Amherst

We evaluate whether open-weight LLMs can assess Computational Thinking (CT) in middle-school students' finite-state game designs against human-coded labels. A single-game analysis shows near-human agreement on CT behavior detection but failure on structural counts, loop detection, and standards mapping while identifying potential problems before a full multi-game evaluation.

FULL SCHEDULE

TUESDAY, OCTOBER 6

Prompting Strategies
Individual Paper Session
11:00 AM – 12:30 PM
Conference Center: Brigade

Presentations:

- **Automatic Prompt Engineering for Generative AI-Based Essay Scoring**
Presenter(s): Yue Huang, Measurement Incorporated
Automatic prompt engineering (APE) offers a systematic, data-informed approach to prompt optimization for generative AI-based automated scoring, yet its application to essay scoring remains underexplored. This study compares theoretically informed APE approaches, their optimized prompts, and scoring performance, providing evidence of APE's potential to advance automated essay scoring.
- **Prompting Strategies for Generative AI Teaching Simulations**
Presenter(s): Jiyun Zu, ETS; Ikkyu Choi, ETS; Jamie Mikeska, ETS; Beata Beigman Klebanov; Michael Suhan, ETS
We study the cognitive fidelity of generative AI-simulated students across different AI models and prompting strategies using an elementary science teaching simulation. Preliminary results highlight the complexity of such simulations, revealing interactions among generative AI model, prompting strategy and targeted student level.
- **Prompting as Measurement: Enhancing Coding for Systematic Reviews with Large Language Models**
Presenter(s): Danielle (Dandan) Kaptur, Pearson; Yue Huang, Measurement Incorporated; Yanhui Guo, University of Illinois Springfield; Xuejun Ji, Medical Council of Canada; The University of British Columbia
Prompting functions as a measurement condition that shapes how LLMs interpret coding tasks. Comparing two prompting approaches using zero-shot and few-shot prompting for GPT-4o-based systematic review coding, we find that the second condition with few-shot prompting produces more conservative yet more precise coding patterns, with implications for reliability and interpretability.
- **Do This, Not That: Can Proscriptive Prompting Reduce LLM Bias?**
Presenter(s): Lauren White, WestEd; Sarah Quesen, WestEd; Michael Chajewski, Pearson
This study extends prior work on structural bias in LLM-generated educational content. Three models (Claude, GPT, Gemini) respond to scaffolded prompts about achievement gaps under two conditions: prompts alone, or prompts with a justice-oriented, bias-reducing rubric. Generated responses are evaluated against the same rubric.

Synthetic Students, Responses & Simulation**Individual Paper Session****11:00 AM – 12:30 PM****Conference Center: Duquense****Presentations:**

- **Multimodal Item Parameter Estimation using Simulated Response Probabilities**

Presenter(s): Christopher Ormerod, College Board; YoungKoung Kim, College Board

We fine-tune a multimodal LLM to replicate students' choice probabilities on image-and-text multiple-choice items across labeled ability levels. By reproducing systematic error patterns, the model implicitly captures 3PL and MCM response curves, letting us accurately estimate held-out item difficulty from its predicted option probabilities.

- **Using AI-Simulated Response Probabilities for Item Difficulty Estimation**

Presenter(s): Brad Bolender, Finetune by Prometric; Sara Vispoel

This paper evaluates an AI-based simulated administration approach for estimating item difficulty before empirical calibration. AI-generated option probabilities for synthetic examinees were converted into item difficulty estimates and compared with operational Rasch difficulty statistics. Results showed meaningful alignment, highlighting simulated administration as a potential tool for earlier item performance evidence.

- **Towards evaluating teacher performance in GenAI teaching simulation of a science discussion**

Presenter(s): Beata Beigman Kelbanov, ETS; Jamie Mikeska, ETS; Mengxuan Zhao, Nesto, Canada; Catherine Flynn, Fidelity Investments; Devon Fetrow, ETS; Shreyashi Halder, ETS; Rutuja Ubale, ETS; Tricia Maxwell, ETS; Michael Suhan, ETS

We investigate systems for automated evaluation and feedback on teacher performance in a GenAI-powered teaching simulation. The findings underscore the importance of a rigorous evaluation of such systems, as we observe poor generalization of models developed using data from a human-puppettered version of the simulation of the same scenario.

- **Using Generative AI to Simulate Item Responses by Skills Insight Ability Band**

Presenter(s): Jianshen (Cassie) Chen, College Board; Chansoon (Danielle) Lee, American Board of Internal Medicine; Kylie Gorney, Michigan State University; Judit Antal, College Board

This study evaluates whether GPT-5.4 Thinking can simulate responses to reading items using Skills Insight ability-band descriptions and item content. Classical and IRT statistics from simulated responses were compared with empirical estimates. Mean IRT b-parameter correlations were moderately strong across conditions, supporting early-stage item evaluation prior to operational field testing.

- **From Static Assignments to Process Evidence: Demonstrating AI-Resilient Simulation-Based Assessment**

Presenter(s): Kumar Srivastava

QLM Studio demonstrates an AI-assisted workflow for converting static assignments into interactive, process-evidence simulations. Learners predict, act, observe, explain, revise, and transfer while the system captures reasoning traces, misconception signals, confidence calibration, and teacher-facing evidence packets to support formative intervention, validity inquiry, and AI-resilient assessment design.

FULL SCHEDULE

TUESDAY, OCTOBER 6

Validating AI in Educational Testing: Challenges and Future Directions

Coordinated Paper Session

11:00 AM – 12:30 PM

Conference Center: King's Garden 5

Session Description:

As generative AI becomes embedded across the testing pipeline, including scoring, item generation, multimodal response evaluation, benchmark construction, and AI-mediated performance itself, validity theory faces pressure at several distinct points. This session brings together five papers, each locating a different fault line exposed by AI integration. The first develops an Evidence-Centered Design heuristic framework for fair, defensible multimodal AI scoring in engineering education. The second argues that generative AI changes the evidentiary basis of response process validity when performance is co-produced through human–AI interaction. The third examines programmatic prompting strategies for LLM-based item generation across content areas, documenting where generated items fail professional review. The fourth proposes a psychometrically-informed framework for evaluating the AI field's own benchmarks. The fifth extends a socio-technical taxonomy of test invalidity to separately diagnose failures in a tool's structure versus its use-plan, empirically testing whether human experts and LLMs classify real AI-testing failures consistently. Together, these papers move from foundational questions about evidence, to applied questions about scoring and generation, to architectural questions about diagnosing failure, offering a structured map of where validity theory must adapt, and where it already has the tools to do so.

Discussant: Daniel McCaffrey (ETS)

Chair: Sergio Araneda (Caveon)

- **Towards Fair and Valid Use of AI to Score Multimodal Constructed Responses**
Presenter(s): Maria Oliveri, Purdue University
- **Rethinking Validity in Educational Assessment when AI Co-Produces Performance**
Presenter(s): Xiaoran Li, Cognality Learning; Tanesia Beverly, AWS
- **Programmatic Tiered Prompting for LLM Generation of ELA & Math Items**
Presenter(s): David Whitcomb, ATLAS, University of Kansas; Marjorie Wine, Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas; Alexander Hoffman, AleDev Research & Consulting
- **Developing a Framework for Psychometrically-Informed Benchmarks for AI in Educational Applications**
Presenter(s): Magdalen Beiting-Parrish, EdAlfy, CUNY Graduate Center
- **Tests as Socio-Technical Artifacts: A Taxonomy for Invalidity Hypotheses**
Presenter(s): Sergio Araneda, Caveon

Validity Arguments & Evidence**Individual Paper Session****11:00 AM – 12:30 PM****Conference Center: Commonwealth 2**

- **Construct Validity of Small-Sample Transformer Scoring Models:**

- **A Mechanistic Interpretability Approach**

- Presenter(s): Michael Hemenway, Pearson; Martha Bellows, Pearson*

- This study applies mechanistic interpretability methods to examine whether transformer-based automated scoring models trained on minimal data (n=64) maintain construct validity comparable to larger-sample models. Preliminary results from 90 DeBERTa-v3-base models largely support construct equivalence across training sizes, with response length emerging as the primary surface feature warranting monitoring.

- **Same Score, Same Meaning? Generative AI, Automated Scoring, and the Validity of Score Interpretations**

- Presenter(s): Christopher Runyon, National Board of Medical Examiners*

Drawing on Kane's scoring inference, we argue that this GAI model development opaqueness is a validity threat, and under some conditions limits the attribution of subject matter expert score meaning to GAI automated scores, even when the scores agree. We distinguish text-proximal from expertise-dependent scoring tasks, walk through an example of scoring a common per020 formance assessment in medical education with analytic and holistic rubrics, and identify the resulting implications for interpretations and use.

- **A Multi-Agent Architecture for Valid, Reliable, and Scalable Skills Assessment**

- Presenter(s): Megan Imundo, WGU Labs; Kjorte Harra, WGU Labs; Lesley Reilly, WGU Labs; Cory Hammon, WGU Labs; Lauren Zito; Catrina Nieser, WGU Labs; Betheny Gross, WGU Labs*

Credentials often fail to capture skills developed outside formal education, limiting access to opportunity. We introduce Current Skills Validation, a multi-agent architecture that decomposes skills assessment into a set of specialized agents, supporting fine-grained adaptivity grounded in learning science. We describe its architecture, foundations, and measurement agenda.

- **GenAI Evidence Hub for Assessment: Methods and Results from 250+ Studies**

- Presenter(s): John Whitmer, Learning Data Insights; Magdalen Beiting-Parrish, EdAlfy, CUNY Graduate Center*

We report findings from a structured literature review of Generative AI (GenAI) empirical research in automated scoring, automated item generation, and formative feedback. Eight trained reviewers used a scoring guide to systematically code studies. We synthesize models used, methodologies, results, and evidence gaps to inform research, practice, and standards development.

- **Source-Text-Conditioned Cloze Generation for Discourse Comprehension Assessment in Intelligent Textbooks**

- Presenter(s): Langdon Holmes; Scott Crossley, Vanderbilt University; Wesley Morris; Seungchan Min, Vanderbilt University*

Cloze exercises offer scalable comprehension assessment, but their validity depends on which words are selected as gaps. We compared three automated methods for cloze exercises generated from summaries within an intelligent textbook platform. Conditioning masked language model predictions on the source text (contextuality-plus) produced higher quality and more source-dependent gaps.

FULL SCHEDULE

TUESDAY, OCTOBER 6

Our Future with AI: Graduate Student Perspectives on a Changing Profession

Sponsor Showcase

11:00 AM – 12:30 PM

Conference Center: Commonwealth 1

Session Description: Artificial intelligence is changing how professionals in assessment, educational measurement, and psychometrics conduct research, develop assessments, analyze data, and make decisions. As AI becomes more integrated into the field, graduate students are preparing to enter a workforce in which both the nature of the work and the skills required to do it are evolving. These changes raise important questions about whether graduate training is keeping pace with emerging expectations and how students themselves view their preparation for the future.

This panel will center the perspectives of four doctoral students from four universities who are preparing for careers in assessment, measurement, psychometrics, and related areas. Panelists will discuss how AI is shaping their coursework, research, professional development, and expectations for future employment. The conversation will also explore what students want and need from graduate programs to feel prepared for an AI-integrated workforce, including technical skills, research competencies, and ethical preparation. The panel will be moderated by a professional who currently oversees psychometric work involving AI-enabled products at an ed tech company and previously spent approximately 20 years in academia developing curriculum, conducting research, and mentoring graduate students in psychometrics and educational research methodology.

Keywords: Artificial intelligence, educational measurement, graduate curriculum, workforce preparation

Moderator: Moderator: John Willse

Discussant: Discussant: Julie Jung (Harvard University), Anna Nasyrova (University of Massachusetts, Amherst), Daniel Oyeniran (The University of Alabama) and Jada Willse (James Madison University)

AI Generation and Estimation of Item Parameters and Skill Blueprints

Individual Paper Session

1:45 PM – 3:15 PM

Conference Center: King's Garden 4

- **A Jurisdictional Scan of Automated Scoring Practices in K-12 Summative Assessment Programs**
Presenter(s): Su Zhang; Jennifer Hove, Education Quality and Accountability Office; Zhimei Gu, Education Quality and Accountability Office
There does not appear to be a survey/scan on Automated Scoring (AS) practices in the US state K-12 summative assessment programs since 2019, and technology perceptions and adoptions have also evolved tremendously. Motivated by these factors, a comprehensive jurisdictional scan was conducted for all states/provinces in the US and Canada.
- **LLM-Based Pairwise Judgment for Math Item Parameter Modeling using Workflows and Agents**
Presenter(s): Suhwa Han, Cambium Assessment; Frank Rijmen, Cambium Assessment
This study evaluates the performance of LLM-based pairwise judgments for recovering banked item parameter scales. We apply a multi-agent framework that constructs a structured pipeline of pairwise comparisons aggregated via the Bradley-Terry-Luce model, and evaluate its performance on a large-scale K-12 summative mathematics assessment dataset.

- LLM as Investigator and Judge in Pairwise Comparisons for Item Parameter Modeling**
Presenter(s): Brian Harrold, University of South Carolina; Vincent Fakiyesi, Engineering Education, college of Engineering University of Georgia; Maria D'Brot, Pearson Assessment; Susan Lottridge, Pearson; Justin Barber, Pearson; Michael Hemenway, Pearson
 This study synthesizes feature-based and pairwise-comparison approaches to item parameter modeling by using an LLM as an investigator of item characteristics and a judge of relative item parameters. The proposed framework aims to improve recovery of item parameter rankings and predictions while providing insight into factors associated with the parameters.
- Validating LLM Rating of Item Features for Explanatory Item Response Model.**
Presenter(s): Ahmed Bediwy, Pearson; Mubarak Mojoyinola
 Four human raters and six large language models evaluated six features of 15 grade-5 mathematics items. We investigate how LLM and human ratings differ in rating these features, and whether LLM-rated features recover human judgment in predicting item difficulty within explanatory item response models, benchmarked against the Rasch model.
- Can LLMs Interpret Psychometric Parameters? Quantifying**
Presenter(s): Kazunori Fukuhara, Stanford University; Benjamin Domingue, Stanford University
 LLM simulation suffers from an “over-knowledge problem”, where models perform too well to represent struggling learners. We compare persona-, IRT-, and CDM-based prompting for student simulation and measure how well methods reflect expected behaviors and quantify this bias. Findings show LLMs follow IRT parameters, yet struggle to simulate real abilities.

AI Feedback and Formative Assessment

Individual Paper Session

1:45 PM – 3:15 PM

Conference Center: Brigade

- Evaluating Feedback Focus and Pedagogical Adaptivity in LLM Feedback on Student Writing**
Presenter(s): Norah Almousa, University of Pittsburgh; Shayan Peyghambari Oskoui, University of Pittsburgh; Raquel Coelho, University of Pittsburgh; Gayle Rogers, University of Pittsburgh; Xiang Li, University of Pittsburgh; Diane Litman
 We examine whether LLMs generate feedback aligned with expert teachers’ practices in feedback focus and adaptivity. We present FeedType, a benchmark of annotated teacher and LLM feedback to evaluate this alignment. Results reveal that while LLMs cover most feedback types, they fail to fully replicate teachers’ feedback distributions and adaptivity.
- Evaluating the Potential of LLM-Generated Feedback to Develop Adaptive Expertise for Revisions**
Presenter(s): Tianwen Li, University of Pittsburgh; Lindsay Clare Matsumura, University of Pittsburgh; Elaine Wang, Rand Corporation; Diane Litman; Richard Correnti, University of Pittsburgh; Zhexiong Liu, University of Pittsburgh
 This study evaluates LLM-generated feedback across three prompt designs (generic, theory-based, and theory-based without a word limit) for its potential to develop seventh-grade students’ adaptive expertise for revisions. The results show that integrating learning theory into prompts can improve feedback quality, but constraining feedback length may sacrifice quality.
- AI-Driven Formative Assessment to Support Growth in Teachers’ Adaptive Expertise for Dialogic**
Presenter(s): Richard Correnti, University of Pittsburgh; Marguerite Walsh, University of Pittsburgh; Tianwen Li, University of Pittsburgh; Lindsay Clare Matsumura, University of Pittsburgh; Diane Litman
 We propose a framework for evaluating Chat-GPT 4.0’s potential to identify lesson segments suited to formative assessment (12 teachers, 35 lessons). Findings indicate that ChatGPT has some potential for selecting segments that can support the pedagogical reasoning processes that develop teachers’ adaptive expertise.

- Source Integration in Writing: Validating a Latent Construct for Formative Feedback**

Presenter(s): Andrew Potter, Yu Tian, Arizona State University; Renu Balyan, SUNY Old Westbury; Maria Goldshtein, Arizona State University; Laura Allen, University of Minnesota; Danielle McNamara, Arizona State University

This study develops and validates an automated, theory-grounded latent measure of source integration from NLP features. Using confirmatory factor analysis across two corpora, the construct aligned with human ratings, generalized across tasks, and improved predictive models. The construct offers an interpretable foundation for integrating LLMs into source-based writing formative feedback.

- Beyond “Three-in-a-Row”: Toward Fairer Adaptive Mastery Decisions in AI-Supported Formative Assessment**

Presenter(s): Chunling Niu, The University of Texas at San Antonio; Kelly Bradley, Marshall University; Brian Waltman, The University of the Incarnate Word; Ashley Love, Southern New Hampshire University; Eugene Kennedy, Louisiana State University

We compare the traditional three-correct-in-a-row mastery rule with an adaptive, measurement-informed mastery model using ASSISTments Skill Builder data. The adaptive model produced more consistent decisions overall, especially early in learning, while rule-based mastery varied across skills, templates, and hint use, suggesting a need for context-sensitive formative decision frameworks.

Advances in AI Scoring and Assessment

Individual Paper Session

1:45 PM – 3:15 PM

Conference Center: Ft. Pitt

- Long-Context Transformer and State-Space Architectures for Data-Efficient Automated Essay Scoring**

Presenter(s): Justin Barber, Pearson; Michael Hemenway, Pearson; Martha Bellows, Pearson; Susan Lottridge, Pearson

We compare long-context architectures—ModernBERT, Longformer, and causal and bidirectional Mamba—for data-efficient automated essay scoring across four training sizes on ASAP-2. Bidirectional Mamba matches the transformers. Deployment readiness depends more on label quantity—rising from 60% to 84% as training grows from 128 to 512 essays—than on backbone choice.

- Automated Justification-Depth Scoring for Adaptive Support in AI-Supported Instructional Tasks**

Presenter(s): Songhee Han, Florida State University; Jueun Shin; Jiyeon Han

This study evaluates machine-learning classifiers for detecting lower justification depth in student responses from AI-supported instructional design tasks. Using 600 human-coded responses and 20 group-aware repeated train-development-test partitions, RoBERTa outperformed TF-IDF logistic regression and Complement Naive Bayes under recall-prioritized thresholding for formative assessment support.

- To Flag or Not to Flag? Detecting Concerning Content in Situational Judgment**

Presenter(s): Susha Suresh, Acuity Insights; Cole Walsh, Acuity Insights; Rodica Ivan, Acuity Insights; Colleen Robb, Acuity Insights

This study evaluates automated approaches for detecting concerning content in open-response SJTs used in higher education admissions. Comparing fine-tuned BERT models with zero-shot and fine-tuned LLMs, we found that fine-tuned BERT achieved the strongest performance despite not receiving the scenario context available to the LLMs

- AI-Enabled Performance Assessment of Communication Skills**

Presenter(s): Lei Liu, ETS; Jiangang Hao, ETS; Enruo Guo, ETS; Jessica Andrews-Todd, ETS; Lin Gu, Educational Testing Service; Amy Adair, ETS; Jonathan Rochkind, ETS; Ou Lydia Liu, ETS

This demonstration showcases an AI-enabled performance assessment prototype of communication skills aligned with a communication learning progression. Attendees will experience how AI-supported interactions, multimodal evidence, and learning progression-guided prompts can support authentic measurement and feedback beyond traditional communication assessments.

FULL SCHEDULE

TUESDAY, OCTOBER 6

Evaluating LLM-Generated Educational Content for Teaching, Learning, and Assessment Coordinated Paper Session

1:45 PM – 3:15 PM

Conference Center: King's Garden 5

Session Description:

This organized session brings together researchers from multiple institutions to examine how the output of large language models (LLMs) can be evaluated across teaching practice, feedback, item development, and mathematics assessment. The four papers move beyond the question of whether LLM outputs appear plausible to ask how validity, fairness, and instructional value can be systematically assessed.

In Presentation 1, Bulut et al. address the “scoring paradox” in AI-generated feedback, comparing score-blind and score-informed approaches through a multi-agent architecture designed to reduce confirmation bias and improve pedagogical quality. In Presentation 2, Choi et al. propose a multidimensional framework for evaluating LLM-simulated students, emphasizing conceptual fidelity, developmental plausibility, interactional realism, instructional usefulness, and fairness. In Presentation 3, Gorgun investigates whether a multi-agent system can serve as an automated judge to detect and revise flaws in AI-generated educational content, such as ambiguity, bias, or content misalignment. In Presentation 4, Arslan et al. evaluate LLM-generated personalized mathematics word problems (MWP) by developing rule-based and LLM-based Natural Language Processing pipelines to determine whether personalized MWPs remain mathematically equivalent to their standard counterparts.

Our discussant, Matthew S. Johnson, will synthesize the findings and examine their implications for addressing methodological challenges in evaluating LLM-generated educational content.

Discussant: Matthew Johnson (ETS Research Institute)

Chair: Burcu Arslan (ETS Research Institute)

- **The Scoring Paradox: Multi-Agent Architectures for Unbiased Feedback Optimization**
Presenter(s): Okan Bulut, University of Alberta; Bin Tan, University of Alberta; Elisabetta Mazzullo, University of Alberta; Cole Walsh, Acuity Insights
- **A Framework for Evaluating LLM-Based Simulated Students to Support Teacher Learning**
Presenter(s): Ikkyu Choi, ETS; Jiyun Zu, ETS; Jamie Mikeska, ETS; Beata Beigman Kelbanov, ETS; Michael Suhan, ETS
- **Can LLMs Detect Item Flaws in the Questions They Have Generated?**
Presenter(s): Guher Gorgun, University of Georgia
- **Automated Evaluation of Mathematical Equivalence Between Personalized and Standard Word Problems**
Presenter(s): Burcu Arslan, ETS Research Institute; Ikkyu Choi, ETS; Jesse Sparks, ETS; Reginald Gooch, ETS; Candace Walkington, Southern Methodist University; Matthew Bernacki, University of North Carolina at Chapel Hill

Fairness, Bias & DIF

Individual Paper Session

1:45 PM – 3:15 PM

Conference Center: Heinz

- Measuring the Vague Meanings of Probability Terms in LLM Scoring**
Presenter(s): Joe Grochowalski, College Board; Claudia Ventura, College Board
 When AI scores an essay “likely proficient,” is there a quantitative basis—and would it produce that phrase from that number? We measured how four large language models interpret and generate probabilistic thresholds in essay scoring. Each maps words to probabilities differently, drifts with context, and often contradicts itself.
- A Multi-Year Investigation of Undergraduates’ Generative AI Fairness Appraisals in Higher Education**
Presenter(s): Sunday Stein, San Diego State University & University of California, San Diego; Victoria Delaney, San Diego State University; Mich Manrique, San Diego State University and University of California, San Diego
 Our work examines the evolution of undergraduate students’ appraisals of genAI fairness in school curriculum across a two-year horizon. We identify and present differences in students’ fairness appraisals and rationales between this academic year and last, delving into underlying sources of change and their implications for higher education policy.
- Effects of Demographic Representations and Model Training Methods on Automated Scoring Engines**
Presenter(s): Yangmeng Xu, Pearson; Martha Bellows, Pearson; Edward Wolfe, University of Iowa
 This study evaluated whether automated scoring engines maintain stable performance when training data composition and training methods vary. We manipulated demographic representation (gender, English language learner, race, student with disabilities) and compared feature-based versus transformer-based models. Results showed performances were stable across subgroup-representation densities and transformer models exhibited greater stability.
- Partial Identification with Multiple Nonlinear Measurements of a Latent Regressor**
Presenter(s): Michelle Yin, Northwestern University; Burhan Ogut, American Institutes for Research
 Multiple AI systems that score the same objects often disagree, so choosing one score biases later analyses. We develop a minimum-distance estimator that combines several AI scores without knowing their scales, and a closed-form interval that bounds the remaining bias from nonlinear disagreement. We validate it on six exposure measures.
- Demonstrating FairGALte: An Agentic LLM Tool for Detecting Construct-Irrelevant Variance**
Presenter(s): Magdalen Beiting-Parrish, EdAlfy, CUNY Graduate Center
 Construct-irrelevant variance—extraneous task elements that unfairly advantage or disadvantage specific examinee groups—critically undermines test validity. This session introduces FairGALte, an open-source, multi-agent LLM-based system that evaluates pilot multiple-choice items for cultural, linguistic, and demographic biases. Attendees will learn to leverage this tool to ensure assessment fairness.

Human-Centered AI Applications: Bridging Learning Science, Assessment, and Ethics

Panel Session

1:45 PM – 3:15 PM

Conference Center: Commonwealth 2

Session Description:

AI is rapidly transforming education, moving beyond novelty and becoming a structural force in teaching, learning, and assessment. This panel convenes five distinguished leaders from industry, academia, and national AI centers, to examine how multiple perspectives allow AI applications to be designed to facilitate learning, ensure valid measurement, and expand educational opportunities. Panelists include the founding dean of the nation's first College of AI, Cyber and Computing, a chief learning officer pioneering AI-driven tutoring, experts in AI ethics and workforce readiness, and advocates for inclusive STEM pathways. Together, they will explore evidence-based approaches to AI design, drawing on learning science for scaffolding and adaptive feedback. The discussion will address the challenges of measuring genuine learning in AI-rich environments, risks and remedies for algorithmic bias, and the evolving role of teachers as collaborators with intelligent systems. It will also confront the tension between policy and practice: as AI innovation and investment accelerate, schools are moving to limit its use, often in the absence of clear evidence of effectiveness. Across topics, the panel will emphasize human-centered AI as a set of design, measurement, and governance choices that determine how responsibility, authority, and interpretability are shared between humans and AI systems.

Moderator: Ou Lydia Liu (ETS)

- **Human-Centered AI Applications: Bridging Learning Science, Assessment, and Ethics**

Presenter(s): Kristen Dicerbo, Khan Academy; Fred Oswald, UC Irvine; Thema Monroe-White, George Mason University; Jinjun Xiong; Xiaoming Zhai, University of Georgia

AI is rapidly transforming education, moving beyond novelty and becoming a structural force in teaching, learning, and assessment. This panel convenes five distinguished leaders from industry, academia, and national AI centers, to examine how multiple perspectives allow AI applications to be designed to facilitate learning, ensure valid measurement, and expand educational opportunities. Panelists include the founding dean of the nation's first College of AI, Cyber and Computing, a chief learning officer pioneering AI-driven tutoring, experts in AI ethics and workforce readiness, and advocates for inclusive STEM pathways. Together, they will explore evidence-based approaches to AI design, drawing on learning science for scaffolding and adaptive feedback. The discussion will address the challenges of measuring genuine learning in AI-rich environments, risks and remedies for algorithmic bias, and the evolving role of teachers as collaborators with intelligent systems. It will also confront the tension between policy and practice: as AI innovation and investment accelerate, schools are moving to limit its use, often in the absence of clear evidence of effectiveness. Across topics, the panel will emphasize human-centered AI as a set of design, measurement, and governance choices that determine how responsibility, authority, and interpretability are shared between humans and AI systems.

FULL SCHEDULE

TUESDAY, OCTOBER 6

Item Difficulty Estimation Individual Paper Session

1:45 PM – 3:15 PM

Conference Center: Smithfield/Birmingham

- **Using LLM Judges' Paired Comparisons to Estimate Mathematics Item Difficulty**

Presenter(s): Lanrong Li, MetaMetrics, Inc.; Mohammed Abulela, MetaMetrics, Inc. and University of Minnesota; Matthew Gushta, MetaMetrics, Inc.

Estimating item difficulty without collecting field testing data has been a long sought-after goal in educational measurement. We prompted large language models (LLMs) to compare mathematics items in pairs to estimate their difficulty. Results showed strong correlations between difficulty based on one LLM judge's paired comparisons and empirical item difficulty.

- **Evaluating Multiple Models for Predicting Item Difficulty in a Principled Assessment Design**

Presenter(s): Alexandra Perez, Cambium Assessment, Inc.; Christina Schneider, Cambium Assessment, Inc.; Sangdon Lim, Cambium Assessment; Garron Gianopoulos, Cambium Assessment; Kang Xue, Cambium Assessment

Item difficulty should, in theory, be predictable from features derived from RPLDs, task characteristics, and linguistic complexity. This study will evaluate the performance of multiple statistical and machine learning models, in estimating item difficulty. We will identify the most influential features contributing to difficulty across grades 3–8 in ELA.

- **Consensus without Accuracy: Investigating LLMs Recovery of Item Difficulty Using Paired Comparisons**

Presenter(s): Michael Chrzan, Center for Educational Data Science and Innovation, University of Maryland; Benjamin Domingue, Stanford University

This study evaluates whether large language models (LLMs) can recover item difficulty estimates through Bradley-Terry modeling from pairwise comparisons of items. Across five Item Response Warehouse datasets and four LLMs, we examine alignment between LLM pairwise-derived difficulty rankings and 1PL IRT parameters, with implications for scalable, AI-assisted item calibration.

- **When One Benchmark Hides Many Truths: Mixture IRT and LLM Difficulty Prediction**

Presenter(s): Enis Dogan, NCES; Burhan Ogut, American Institutes for Research

We examine whether LLM predicted item difficulty aligns differentially with class specific empirical estimates from mixture Rasch models. Using responses ($N \approx 14,501$) to a 29 item math exam and pairwise judgments from gemma4:26b aggregated via Bradley-Terry, we find LLMs track high ability examinees best, revealing limits of single benchmark validation and we recommend class specific validation practices.

- **CAP: Cross-Ability Preferences to Align Student Simulators for Item Difficulty Prediction**

Presenter(s): Nigel Fernandez, University of Massachusetts Amherst; Alexander Scarlatos, University of Massachusetts Amherst; Christopher Ormerod, College Board; Andrew Lan, University of Massachusetts, Amherst

Predicting open-ended item difficulty from few labeled examples is challenging. We present CAP (Cross-Ability Preferences), which aligns an LLM-based student simulator with direct preference optimization (DPO) over pairs contrasting students of different IRT abilities. On programming items, CAP outperforms the state-of-the-art simulator SMART at 1B and performs comparably at 3B.

Multimodal and Chatbot-Based Tutoring**Individual Paper Session****1:45 PM – 3:15 PM****Conference Center: Duquense**

- **A Cognitive Lab Study of Student AI Chatbot Use in Online Coursework**

Presenter(s): Jinah Choi, Edmentum; Sonya Powers; Farzan KarimiMalekabadi, Edmentum; Michelle Barrett, Edmentum

This cognitive lab study examines how high school students interact with an AI chatbot embedded in a K-12 online learning platform. Using chat logs, facilitator observations, post-session reflections, and surveys, the study identifies authentic student interaction patterns to inform chatbot design, safeguards, and provide response process validity evidence.

- **Developing and Validating an Automatic Scoring Model for Chatbot-Based Conversational Speech**

Presenter(s): Danwei Cai, Duolingo; Audrey Kittredge; Ben Naismith, Duolingo; Xiangying Jiang, Duolingo; Kevin Yancey, Duolingo

This paper describes the development and validation of an automated scoring model for open-ended chatbot-based conversational speech among English learners in the Duolingo learning app. The model strongly predicted human ratings, produced reliable scores, and showed concurrent validity with Duolingo English Test speaking items, demonstrating stealth proficiency assessment at scale.

- **A Multimodal AI Analysis of Teacher Praise in Online Tutoring**

Presenter(s): Jiseung Yoo

This study develops a multimodal AI framework to measure teacher praise quality by integrating linguistic content and acoustic delivery in online tutoring sessions. Large language models identified praise and specific praise, while a WavLM-based classifier detected enthusiastic delivery. Results showed that praise quality mattered more than praise quantity.

- **Breaking the Learning System Wall Using Multimodal Tutoring Transcriptions**

Presenter(s): Danielle Thomas, Carnegie Mellon University; Marie Cynthia Abijuru Kamikazi, Carnegie Mellon University; Conrad Borchers, Vanderbilt University; Ashish Gurung, Carnegie Mellon University; Ishan Miglani, Carnegie Mellon University; Shivang Gupta, Carnegie Mellon University; Zachary Levonian, Renaissance Philanthropy; Ken Koedinger, Carnegie Mellon University

Few methods exist for generalizing student learning across learning platforms. To break the “learning system wall,” we demonstrate a method of generating multimodal transcriptions of remote tutoring where students are engaging with a human tutor and MATHia platform. We successfully extract structured logs and compare to MATHia as ground truth.

- **Framing validity arguments from student multimodal collaborations with generative artificial intelligence**

Presenter(s): Benjamin Pierce, Institute for Intelligent Systems; John Sabatini, Institute for Intelligent Systems, University of Memphis

When students collaborate with generative AI to create multimodal artifacts, what can we validly infer about their competence? Using process data from a large study of AI-supported scenario-based assessment, we draw on established measurement principles to show how multimodal AI collaboration can be treated as construct-relevant.

**Operational Trustworthiness in Automated Scoring: Validity, Bias, Error, Stability, Cost
Coordinated Paper Session****1:45 PM – 3:15 PM****Conference Center: Commonwealth 1****Session Description:**

This coordinated paper session examines how automated scoring systems can be evaluated, monitored, and defended under operational conditions. The papers move beyond overall agreement as the primary evidence for trustworthiness and instead examine validity checks, conditional bias, error diagnosis, model stability, and cost-aware sufficiency. One paper evaluates small generative models for detecting invalid and fluent off-topic essay responses before ordinary scoring occurs. A second investigates how automated, human, and expert/backread scores differ by task type, score level, and reused prompt, showing how aggregate agreement can hide localized monitoring risks. A third reframes compact AES distillation as a stability-bounded inference problem, examining perturbation sensitivity, out-of-distribution shifts, and subgroup uncertainty. A fourth uses unsupervised audio-embedding methods to identify interpretable categories of speech-scoring errors in early-literacy assessment. A fifth proposes cost per valid score as a measurement-informed unit for comparing scoring designs only after validity, reliability, fairness, and operational sufficiency constraints are declared. Together, the session treats trustworthiness as an evidence system: valid scores require not only accurate models, but defensible decisions, interpretable failures, ongoing monitoring, and sustainable operational design.

*Discussant: Susan Lottridge (Pearson)**Chair: Edward Wolfe (University of Iowa)*

- **Beyond Agreement: Conditional Bias and Operational Monitoring of Automated Scoring Across K–12**
Presenter(s): Jing Ma, University of Iowa; Edward Wolfe, University of Iowa; Anthony Fina, University of Iowa
- **Detecting Invalid Responses in Automated Essay Scoring Using Large Language Models**
Presenter(s): YoungKoung Kim, College Board; Christopher Ormerod, College Board
- **Compact AES Distillation Within a Bounded Stability Regime**
Presenter(s): Yi Gui, Measurement Inc
- **Making Sense of Speech Scoring Errors with Unsupervised Learning**
Presenter(s): Zachary Schultz, Cambium Learning Group, Inc.; Kai North, Cambium Learning Group, Inc.; Gitit Kehat, Cambium Assessment; Alexander Kwako, Cambium Assessment; Debbie Dugdale, Cambium Assessment, Inc.
- **From Efficiency to Sufficiency: A Measurement-Informed Framework for Cost-Aware Automated Essay Scoring**

FULL SCHEDULE

TUESDAY, OCTOBER 6

Recent AI Applications at Measurement Incorporated (MI)
Coordinated Paper Session
1:45 PM – 3:15 PM
Conference Center: Grand Ballroom 1

Session Description:

Artificial intelligence (AI) is rapidly transforming educational measurement, assessment, and learning analytics, creating opportunities to enhance scoring, feedback, psychometric modeling, and construct representation. This symposium gives an overview of the recent AI applications at Measurement Incorporated (MI). It examines complementary approaches that integrate AI with established measurement frameworks while preserving validity, interpretability, and fairness. The first study explores the use of generative AI as a feature-augmentation tool for automated essay scoring (AES) with model comparisons among traditional feature-based, AI-derived construct features, and hybrid approaches. The second study develops an agreement-conditioned framework for interpreting paired standardized mean differences in human–automated scoring evaluations, providing practical guidance for validation and monitoring of AI-assisted scoring systems. The third study investigates an AI-assisted formative feedback system for rater training and examines its effects on qualification success, scoring accuracy, agreement, and operational scoring quality. The fourth study compares Deep Item Response Theory (DeepIRT) with traditional IRT models, evaluating predictive performance, latent representations, and interpretability. The fifth study demonstrates the discourse-type-conditioned interpretations for AES and AI-driven writing feedback. Collectively, these studies highlight the promise of AI-enhanced assessment systems while emphasizing the continued importance of psychometric rigor, explainability, fairness, and construct validity in next-generation educational measurement.

Discussant: Joshua Wilson (University of Delaware)

Chair: Duanli Yan (Measurement Incorporated)

- **Rubric-Aligned Generative-AI Features as Supplementary Predictors in Feature-Based Automated Essay Scoring**

Presenter(s): Yue Huang, Measurement Incorporated; Duanli Yan, Measurement Incorporated; Corey Palermo, Measurement Incorporated

Feature-based automated essay scoring (AES) relies on explicitly defined linguistic features, which are essential but may represent rhetorical qualities like organization and evidence incompletely. Rubric-aligned generative-AI features could fill this gap, yet their incremental value over established linguistic features remains untested. Using 4,572 Grade 7 argumentative essays scored on three traits—Organization/Purpose, Evidence/Elaboration, and Conventions—we compared random forest models using 60 linguistic features, 42 generative-AI features, or a combined feature set. We found that generative-AI features correlated moderately with human trait scores ($|r| = .29$ to $.64$) and overlapped only partially with the linguistic features. On a held-out test set, combined models improved quadratic weighted kappa (QWK) by $.011$ to $.019$ over linguistic-only models. Paired bootstrap intervals excluded zero for Organization/Purpose QWK, as well as Conventions exact agreement and mean absolute error. Our findings support the integration of rubric-aligned generative-AI features alongside linguistic features to meaningfully improve trait-based AES.

- **Standardized Mean Differences Under Response-Level Dependence: Implications for AI-assisted Automated Scoring**

- **AI-Assisted Rater Training Modules: Effects on Qualification and Scoring Quality**

Presenter(s): Jie Chen, Measurement Incorporated; Duanli Yan, Measurement Incorporated; Yong He

- **Beyond Traditional IRT: Exploring DeepIRT for Next-Generation Educational Measurement**

Presenter(s): Mark Beck, Measurement Incorporated; Duanli Yan, Measurement Incorporated; Yong He

- **Latent Factors of Discourse Type and Effectiveness in PERSUADE 2.0 for AES**

Presenter(s): Yi Gui, Measurement Incorporated (MI)

AI for Alignment and Construct Validity

Individual Paper Session

3:45 PM – 5:15 PM

Conference Center: Brigade

- **Using Natural Language Processing to Explore Alignment Between Skill Taxonomies**
Presenter(s): Teresa Ober, ETS; Michael Flor, ETS
 We examine alignment at both node and taxonomy levels across four major skill taxonomies. Using semantic similarity and cluster analysis, results show potentially meaningful though uneven overlap: some skill nodes are semantically similar, and some taxonomies show stronger interleaving than others. Findings show how AI-based approaches could identify construct overlap.
- **Pre-Data Embedding Diagnostics for Construct Validity in Multi-Domain Instruments**
Presenter(s): Youngmi Cho, Riverside Insights; Tong Wu; Hyeonjoo Oh, Riverside Insights
 Centroid margin, a relative embedding metric capturing an item's proximity to its intended domain relative to competing domains, is evaluated as a pre-data tool for sharpening construct boundaries. Across four samples, it improved model fit and discriminant validity; cosine similarity counterintuitively reduced discriminant validity in the most overlapping domains.
- **Auxiliary Information for Semantic Clustering of Assessment Items**
Presenter(s): Josiah Hunsberger, James Madison University; Aquia Richburg, National Commission on Certification of Physician Assistants; Marcus Walker, National Commission on Certification of Physician Assistants
 This study identifies a selected transformer-based clustering model for operational medical assessment items that balances within-family proximity with cluster separation. Operational analyses showed supplementary alignment with blueprint structure, flagged isolated content areas that may need additional item coverage, and identified dispersed topics for SME review.
- **Supporting Distractor Quality Review Through Interpretable Semantic and Lexical Diagnostics**
Presenter(s): Michelle Banawan, Arizona State University; Shubham Chakraborty, Arizona State University; Linh Huynh, Arizona State University; Katerina Sabino, Arizona State University; Tracy Arner, Arizona State University; Danielle McNamara, Arizona State University
 Multiple-choice distractor quality is often evaluated through qualitative review. We present MCQ-Diag, a measurement-informed diagnostic tool that provides option-level semantic and lexical indicators for distractor inspection, revision, and corpus-level analysis. Expert validation suggests the indicators align with human judgments of distractor plausibility and linguistic quality.
- **Estimating, Calibrating, and Explaining Domain Proficiency from Score Trajectories**
Presenter(s): Keith Holliday, Curriculum Associates
 This study evaluates a dynamic mathematics proficiency-estimation system that combines machine learning prediction, post-hoc calibration, and LLM-generated explanation. Using longitudinal K–8 assessment histories, it examines SEM alignment, conditional bias across relative grade-level bands, and whether structured LLM rationales make score estimates more interpretable for teachers.

AI for Mathematics: Scoring & Item Development

Individual Paper Session

3:45 PM – 5:15 PM

Conference Center: Smithfield/Birmingham

Specialist vs. Generalist: Comparing MathBERT and DeBERTa for Scoring Math-Explanation Items

Presenter(s): Gregory Jacobs, Pearson; Ahmed Bediwy, Pearson; Martha Bellows, Pearson

We fine-tune a math-specialized pre-trained transformer (MathBERT) and a larger general-purpose pre-trained transformer (DeBERTa-v3-large) to score 41 math-explanation prompts. Both reach human-like agreement on the majority of prompts (Avg. QWK $\approx$ 0.93). MathBERT commonly matches DeBERTa at roughly 40% fewer parameters, suggesting domain pretraining drives efficient scoring for math-explanation items.

Is Hyperparameter Tuning Worth the Compute? Efficiently Fine-Tuned Transformers for Math Scoring

Presenter(s): Gregory Jacobs, Pearson; Ahmed Bediwy, Pearson; Martha Bellows, Pearson

For automated scoring of math-explanation items, does hyperparameter tuning justify its cost? Across 41 prompts, tuning two pre-trained transformers—MathBERT and DeBERTa-v3-large—overwhelmingly yielded gains no larger than fold-to-fold noise. We propose fixed model-specific defaults and reinvesting compute in cut-point optimization, reducing tuning cost roughly tenfold with negligible accuracy loss.

- **Automatic Pre-Testing of Mathematics Items: predicting difficulty parameter with machine learning methods**

Presenter(s): Joao Freire, CAEd/UFJF; Patrick Carvalho, CAEd/UFJF; Luan Barbosa, CAEd/UFJF; Emerson Souza, CAEd/UFJF

This study evaluates machine learning methods for predicting IRT difficulty parameters for mathematics items using text, metadata, and specialized expression tokenization. Across Brazilian basic education datasets, Random Forest models achieved competitive accuracy and preserved downstream proficiency classifications, indicating that automated pretesting can support scalable calibration in large-scale operational assessment systems.

- **Ai-Made Test Construction In Mathematics: A Comparison with Teacher-Made Tests**

Presenter(s): Thuy Dieu Le, University of Massachusetts, Amherst; Thi Nguyet Ha Le, Hanoi National University of Education; Nhat Huy Nguyen, Hanoi National University of Education

We evaluate a human-in-the-loop LLM workflow generating a complete Grade 10 statistics test aligned to Vietnam's curriculum, compared with a teacher-made form via expert content-validity review and student data (N = 190; McDonald's omega, TOST). AI items lose content relevance across sections yet form a more coherent score scale.

- **Using XGBoost to Construct a Vertically Skill Difficulty Scale**

Presenter(s): John Bielinski, Renaissance

An XGBoost prediction model was fitted to online skill practice data for a collection of 1800+ skills to construct a vertically aligned skill difficulty scale spanning kindergarten through 8th grade math. The results were validated through association with an independently derived Rasch vertical scale.

FULL SCHEDULE

TUESDAY, OCTOBER 6

AI for Measurement in Educational Learning and Assessment (upcoming book)

Coordinated Paper Session

3:45 PM – 5:15 PM

Conference Center: Commonwealth 1

Session Description:

Artificial Intelligence in educational learning and assessment captures a pivotal moment in education: the rise of AI as both a tool and a disruptor. No longer confined to efficiency or automation, AI now shapes how learners think and self-regulate, how educators teach and adapt, and how assessment defines knowledge and skill in a globalized world. Currently, education is experiencing the biggest disruption since the digital revolution. AI is not hypothetical anymore; it is embedded in classrooms, exams, tutoring platforms, policymaking. Few collections bring together psychometrics, cognitive science, classroom practice, AI engineering, and ethics in one place. Most publications cover just one of these angles. This coordinated session includes five chapters from the upcoming volume *AI for Measurement in Educational Learning and Assessment*, which takes an interdisciplinary approach. It explores the impact of the new digital affordances: how we define fairness, creativity, equity, and even what it means to “know” something in an AI-rich world. That has resonance for researchers, policymakers, and practitioners alike. It argues that AI is not merely an add-on to existing systems but a transformative force that challenges long-held assumptions about constructs, evidence, and fairness. The presentations capture a moment when the field is being redefined.

Discussant: Jiangang Hao (ETS)

Chair: Duanli Yan (Measurement Incorporated)

- **Introduction to AI for Measurement**

Presenter(s): Alina von Davier, Duolingo; Duanli Yan, Measurement Incorporated

- **Human agency before and after Generative Artificial Intelligence in education, assessment, measurement**

Presenter(s): Andrew Ho, Harvard University

- **The Evolution of a Generative AI-Powered Tutor**

Presenter(s): Kristen Dicerbo, Khan Academy; Kelli Hill, Khan Academy

- **Automated essay scoring in the age of AI**

Presenter(s): Corey Palermo, Measurement Incorporated; Duanli Yan, Measurement Incorporated; Yong He

- **Large Language Models for Multilingual Scoring in TIMSS and PIRLS**

Presenter(s): Matthias von Davier, Boston College; Ummugul Bezirhan, Boston College

Building Responsible AI Practice in Language Assessment: Anticipating What's Next **Panel Session**

3:45 PM – 5:15 PM

Conference Center: King's Garden 5

Session Description:

Language assessment is grounded in measurement science and professional ethics. As AI reshapes test development, scoring, and security, this panel examines responsible AI infrastructure, tensions like scalability versus interpretability, and emerging issues such as agentic AI. Voices from language assessment, educational assessment, and AI in education explore how our fields shape AI use.

Moderator: Jill Burstein (Duolingo)

- **Building Responsible AI Practice in Language Assessment: Anticipating What's Next**

Presenter(s): Jill Burstein, Duolingo; Alina von Davier, Duolingo; Ikkyu Choi, ETS; Sarah Hughes, Pearson; Xun Yan, University of Illinois, Urbana-Champaign; Jing Xu, Cambridge University Press & Assessment; Mukhta Ananda, Gemini, Google DeepMind

Comparing and Validating Human Scoring and Rating with Machine Scoring **Individual Paper Session**

3:45 PM – 5:15 PM

Conference Center: King's Garden 4

- **Assessing the Reliability and Construct Representation of LLM-based Language Proficiency Scores**

Presenter(s): Langdon Holmes; Joon Choi, The Learning Agency; Scott Crossley, Vanderbilt University; Wesley Morris

We used confirmatory factor analysis to assess the reliability and construct representation of an LLM-based measurement instrument of language proficiency. LLMs were at least as reliable as human raters and loaded onto the same underlying factor, though analyses indicated a less than perfect alignment between LLM and human

- **Interpretable Humans, Alien LLMs: Expert Analysis of Latent Structures in Assessment Responses**

Presenter(s): Alona Strugatski; Licol Zeinfeld, Weizmann Institute of Science; Jason Cooper, Weizmann Institute of Science; Shelley Rap, Weizmann Institute of Science; Gil Schwartz, Weizmann Institute of Science; Giora Alexandron, Weizmann Institute of Science

Applying Exploratory Factor Analysis to human and LLM responses on chemistry and quantitative reasoning assessments, we asked domain experts to blindly interpret emergent factors. Experts decoded human factors successfully but could not interpret most LLM factors, suggesting LLMs rely on statistically opaque mechanisms distinct from human cognitive constructs.

- **Validating Large Language Models as Anchor Judges for Cross-National Creativity Comparisons**

Presenter(s): Ricardo Primi, Universidade São Francisco

Cross-national comparison of PISA 2022 creative-thinking scores is limited by missing anchor ratings linking countries' rater pools. We tested a large language model (GPT-5.1) as an anchor judge across Brazil and Portugal. LLM-human agreement matched human-human agreement, supporting LLMs for valid cross-cultural linking via Many-Facet Rasch modeling

- Exploratory Bayesian Multidimensional IRT with Rater Effects for Scoring Generative AI Usage**
Presenter(s): Leon Camus, DIPF | Leibniz Institute for Research and Information in Education; Sebastian Gombert, Leibniz Institute for Research and Information in Education; Fabíola Ribeiro, DIPF | Leibniz-Institute for Research and Information in Education; Gianluca Romano, DIPF | Leibniz-Institute for Research and Information in Education; Carmen Köhler, DIPF; Hendrik Drachler, DIPF | Leibniz-Institute for Research and Information in Education, Computer Science Department, Goethe University
 Students' generative-AI use is a profile of distinct behaviors, not one skill, yet scoring collapses it onto one axis. We develop a multidimensional Bayesian IRT model with rater effects, identified post-hoc by rotation, recovering three dimensions: prompting effort, content delegation, AI-delivered citations. Domain familiarity shifts students from delegation toward engagement.
- Can an LLM Out-Search Hyperparameter Optimization for Automated Essay Scoring?**
Presenter(s): Julius Frost
 We benchmark an LLM-in-the-loop that hypothesizes, edits one training file, and keeps or discards by a metric against four classical hyperparameter optimizers, all wallclock-matched on ASAP-AES. The LLM wins screening via an off-grid ordinal-loss move the optimizers cannot express; a full-fidelity held-out transfer check is underway.

Item Difficulty & Parameter Prediction

Individual Paper Session

3:45 PM – 5:15 PM

Conference Center: Duquense

- Predicting IRT Parameters for Passage-Based Reading Items Using NLP**
Presenter(s): Ruitao Liu, ACT
 This study predicts IRT item difficulty and discrimination for passage-based reading comprehension items using lexical, syntactic, and semantic NLP features. Modeling interactions among passages, stems, and options, LightGBM with semantic embeddings best predicts difficulty ($r = 0.594$), while discrimination remains harder to recover from text alone.
- Predicting Item Statistics Using LLM-Derived Difficulty Ratings and Response-Option Probability Features**
Presenter(s): Nate Smith, American Board of Internal Medicine
 This study examines whether interpretable LLM-derived features can predict difficulty or statistical acceptability for medical certification items. Models reduced prediction error ($R^2 = 0.381$) and predicted p-values strongly ($r = 0.624$). Ability persona prompting rarely changed response accuracy (~90–98%) but produced useful probability mass reallocations across response options.
- Anchored Bradley-Terry Calibration Using LLM Comparative Judgments**
Presenter(s): Ummugul Bezirhan, Boston College; Matthias von Davier, Boston College
 AI-based comparative judgment was evaluated as a tool for early item difficulty estimation. Three AI judges compared new TIMSS Grade 4 mathematics items with calibrated anchors, with rankings analyzed using a fixed-anchor Bradley-Terry model. Results show meaningful difficulty signals, supporting scalable supplementary use while highlighting anchor coverage and comparison-network design.
- Funnel Plot Analysis of Examinee Performance in Longitudinal Assessment**
Presenter(s): Aquia Richburg, National Commission on Certification of Physician Assistants; Marcus Walker, National Commission on Certification of Physician Assistants
 We identify content areas where examinees underperform in a longitudinal assessment for a high-stakes medical licensure exam. Using modified funnel plots with a moving baseline to account for item difficulty, we rank topics by relative performance. Preliminary results suggest factors beyond item difficulty, informing future analyses and educational interventions.

- **Measuring and Modifying Text Complexity: A Multi-Model LLM Evaluation**

Presenter(s): Zack Feldberg, MetaMetrics, Inc.; Mohammed Abulela, MetaMetrics, Inc. and University of Minnesota; Laura Faircloth, MetaMetrics, Inc.; Justin Taylor, MetaMetrics, Inc.; Matthew Gushta, MetaMetrics, Inc.

We evaluate how accurately LLMs estimate and modify text complexity within the Lexile Framework. Using 600 controlled passages across Lexile bins, text lengths, and text types, we benchmark multiple model families and inference modes, quantifying accuracy, bias, replicate stability, directional asymmetry, and content fidelity against the Lexile Text Analyzer.

To Err is not only Human: Human/LLM Collaboration in Assessment

Panel Session

3:45 PM – 5:15 PM

Conference Center: Commonwealth 2

Session Description:

Large language models (LLMs) offer powerful capabilities to support learning, assessment, and research. As examples, an assessment designer might use an LLM to support automated item generation; a teacher might brainstorm with an LLM while building a classroom lesson; a researcher or computer engineer might use LLMs to conduct an evaluation study; a psychometrician might use an LLM to assist with identifying an explanatory IRT model; or a student might use an LLM-based tutor. While they have capacity to synthesize large volumes of data, LLMs also hallucinate, lack true reasoning capabilities, and can produce inconsistent or biased outputs. While limited in their ability to synthesize massive datasets, humans can reason systematically and attend to nuance, fairness, and tone. Because they have complementary strengths and weaknesses, humans and LLMs can be ideal partners when they work collaboratively and iteratively. The panelists will share real-world experiences and lessons learned by professionals in different roles, and highlight practical strategies for leveraging LLMs while maintaining human oversight. Discussions will address prompt engineering, evaluation protocols, bias mitigation, and integration into evidence-centered frameworks. By examining both successes and failures, we aim to advance principled approaches to human-AI partnership that enhance learning outcomes without compromising measurement integrity.

Moderator: Edith Graf (Proof-of-Concept Learning LLC)

- **To Err is not only Human: Human/LLM Collaboration in Assessment**

Presenter(s): Edith Graf, Proof-of-Concept Learning LLC; Manqian Liao, Duolingo; Maria Oliveri, Purdue University; Emily Proctor, College of Engineering, Department of Engineering Education, Purdue University; Alex Rudniy, Drew University

Transforming Assessment Data into Educational Action: Human–AI Collaboration in Educational Measurement

Coordinated Paper Session

3:45 PM – 5:15 PM

Conference Center: Grand Ballroom 1

Session Description:

Artificial intelligence (AI) is rapidly transforming educational measurement, creating new opportunities to enhance design, operational processes, and data analysis. However, researchers and practitioners are also encountering challenges regarding reliability, validity, interpretability, accountability, and human oversight. As AI tools become more integrated into assessment systems, there is a growing need for frameworks and applications that balance innovation with measurement rigor and trustworthiness.

This session examines how human–AI collaboration can help transform assessment data into meaningful educational insights and actions while preserving core measurement principles. The session begins with a conceptual framework for the responsible, human-centered implementation of AI applications in educational measurement. The second paper introduces a novel time-embedded n-gram approach for extracting temporally meaningful patterns from process data to better understand cognitive strategies in a numeracy task. The third paper presents a human–machine learning framework that uses semantic embeddings and recommendation systems to support large-scale qualitative

coding. The final paper introduces a scalable agentic AI framework that combines large language models, human-designed rubrics, and expert examples to generate formative data insights from assessment data.

These papers collectively illustrate how human-AI collaboration can generate richer evidence, improve operational processes, and support educational decision-making while maintaining rigor, transparency, accountability, and trust.

- **Discussant: Paul Kleinhert (Hawai'i State Department of Education)**
Chair: Judy Tang (Westat)
- **Human-AI Collaboration in Educational Measurement: Transforming Assessment Data into Educational Action**
Presenter(s): Laura Egan, Westat; Judy Tang, Westat
- **Understanding Spatial Cognitive Process Using Time-Embedded N-Grams Model with Machine Learning**
Presenter(s): Qiwei He, Georgetown University; Yiming Chen, University of Minnesota; Elizabeth Tighe, Georgia State University; Gal Kaldes, Georgia State University
- **Human-Machine Learning for Large-Scale Qualitative Coding in Educational Measurement**
Presenter(s): Judy Tang, Westat; Tom Krenzke, Westat
- **Scaling Human-AI Collaboration: Translating Multi-Source Assessment Data into Formative Learner Profiles**
Presenter(s): Hongwen Guo, ETS Research Institute; Matthew Johnson, ETS Research Institute; Luis Saldivia, ETS; Michelle Worthington, ETS; Jeremy Lee, ETS; Kadriye Ercikan, ETS Research Institute

Voices from the Field: How AI is affecting classroom assessment?

Panel Session

3:45 PM – 5:15 PM

Conference Center: Ft. Pitt

Session Description:

Generative AI is reshaping classroom assessment by challenging evidence of student thinking, task validity, and authorship. This panel examines how educators redesign tasks, embed formative assessment processes, and make instructional decisions in AI-rich environments. Through moderated dialogue, participants gain practical strategies for aligning assessment practices in an AI world.

Moderator: Bryan Drost (NCOESC)

- **Voices from the Field: How AI is affecting classroom assessment?**
Presenter(s): Bryan Drost, NCOESC; Caitlin Fine, Metropolitan State University of Denver; Char Shyrock, Leadership Consultant; Laine Bradshaw, Pearson; Mike Maksimchuk, Michigan Virtual; Jennifer Ivers
Abstract:

FULL SCHEDULE

TUESDAY, OCTOBER 6

Poster Session & Reception

Poster Reception

5:15 PM – 6:30 PM

Conference Center: King's Garden 1-3

- **When is LLM Confidence Valid Evidence in Educational Measurement?**

Presenter(s): Lingbo Tong, University of Wisconsin–Madison; Jee-Seon Kim, University of Wisconsin–Madison

This study introduces text supportiveness, a response-level construct capturing whether open-ended survey text reinforces, attenuates, or contradicts closed-ended ratings. We evaluate LLM uncertainty quantification methods for probabilistically estimating text supportiveness and propose validity criteria to empirically assess when LLM-derived evidence supports response-process validity in mixed-format survey measurement contexts.

- **Student Perceptions of AI-Assisted Fashion Design Research**

Presenter(s): Tai Wai David Yeung, Baylor University; Jay Yoo, Baylor University

This study examines generative AI use in fashion design education to create mood boards. The learning outcome suggested that students valued AI for ideation during initial design development, but also preferred manual approaches to foster creative personal expression. Results support the benefit of AI integration within studio-based fashion design learning.

- **Generating and Reviewing Mathematics Item Families With AI**

Presenter(s): Andrea Hebert, Cognia

This exploratory study investigates AI-supported workflows for mathematics item development and quality review. Using a released Grade 10 mathematics assessment form, large language models generated parallel item families and anticipated psychometric characteristics. Findings suggest AI may support construct-preserving item development while flagging potential semantic drift and review concerns.

- **TRACE: Automated Multi-Granularity Analysis of Text Revision**

Presenter(s): Yu Tian, Arizona State University

This proposal introduces TRACE, an automated framework for analyzing textual revisions across drafts. TRACE aligns initial and revised texts and infers interpretable revision operations such as insert, delete, replace, and move. TRACE supports AI-powered revision analytics, writing assessment, automated feedback, and studies of human-AI writing collaboration.

- **Comparing θ Representation Strategies for Reconstructing Item Characteristic Curves**

Presenter(s): Sungjin Nam, ACT Education Corp.

We examine how different strategies for representing student skill (θ) affect IRT-3PL item parameter recovery from reconstructed item characteristic curves. Our approaches, using signed decimal or scientific notation, yield more accurate parameter recovery than the baseline approach relying on a fixed set of natural-language ability descriptors, while also offering greater scalability.

- **Hierarchical Attention Network Architecture for Automated Response Time Prediction**

Presenter(s): Wei Schneider, Cambium Assessment

This paper applies the Hierarchical Attention Network (HAN) to item response time prediction using minimal item metadata. HAN improves deep learning model explainability by identifying both words within sentence context and sentences in an item with long response time, bridging the gap between complex algorithmic performance and actionable instructional feedback.

- **Measuring AI-Related Constructs in Education: A Psychometric-Framework for Validating AI Attitude Scale**

Presenter(s): Obed Boateng, Virginia Commonwealth University

As AI integrates into education, measuring AI-related psychological constructs becomes essential to validity, fairness, and policy. Using HMS 2024–2025 data (N = 930), I developed and validated a three-factor AI Attitudes Scale using sequential EFA–CFA, offering a replicable framework for measuring AI-related constructs, contributing a replicable methodological framework for measuring AI-related construct.

- **Calibrated, Interpretable Automated Scoring with LLM Likelihoods**

Presenter(s): S. Thomas Christie, Renaissance Philanthropy; Markus Hauru, Renaissance Philanthropy; Anna Rafferty, Carleton College

Paradigms like IRT use measurement equations to model student behavior probabilistically. We investigate an analogous LLM-based approach, the noisy channel model, to compute and compare the conditional likelihoods of student writing and produce auditable token-level inferences about student skills. We compare against other LLM-based methods on accuracy, calibration, and interpretability.

- **Fine-tuning DeBERTa-v3 to Automate Spatial Language Classification**

Presenter(s): Sam Agnoli, PhD Student; Qingzhou Shi, Northwestern University; David Uttal, Northwestern University

To automate spatial language classification, we fine-tuned DeBERTa-v3, a transformer-based language model, using a 33,158 word dataset following a 70:15:15 train-test-validation split. The model performed well in a held-out test comparing performance to human coders (Cohen's kappa = .86), outperforming a multi-agent foundation model (GPT4-o; Cohen's kappa = .66).

- **Validating AI-Simulated Panelists for Standard Setting**

Presenter(s): Jonathan Henriques, James Madison University; Brian Leventhal, James Madison University; Jada Willse, James Madison University; Brian French, Washington State University

We develop and apply a validation protocol for evaluating AI-simulated standard setting panelists. Using a case study, we examine the capability of prompt-based AI panelists to approximate intended stakeholder perspectives, provide defensible rationales, and respond consistently across replications. Findings will inform responsible uses and limitations of genAI-supported standard setting.

- **Toward Self-Improving Automatic Item Generation: Validating LLM Evaluators with Production Feedback**

Presenter(s): Jenny Han, Khan Academy; Corey Kollbocker, Khan Academy; Tyler Burleigh, Khan Academy; Kavitha Gopalakrishnan, Google; Hasibe Goculu, Google; Emilia Palaghita, Google

A scalable way to measure and improve automatic item generation (AIG) at scale is to build expert-validated LLM evaluators, though expert-labeled datasets are still costly and narrowly scoped. This study proposes combining small expert-labeled datasets with large-scale datasets from production user feedback to build LLM evaluators for an AIG system.

- **Identifying Communication Profiles from Physician-Patient Conversations Using Large Language Models**

Presenter(s): Reyhaneh Hosseinpourkhoshkbari, University of Texas at Dallas; Yash Jain, University of Texas at Dallas; Richard Golden, University of Texas at Dallas

Human evaluators and a Large Language Model applied the widely used Kalamazoo assessment framework to evaluate doctor communication skills for 213 human simulated patient-physician respiratory case interviews. Good overall agreement between the human evaluators and the LLM was obtained but agreement measure performance varied considerably across specific assessment dimensions.

- **Scaffolding AI Use in Introduction to Psychology Research Assignments**

Presenter(s): Jennifer Boyd, University of Tennessee

This work-in-progress examines an AI-integrated instructional model that scaffolds undergraduate students' use of generative AI during research activities in Introduction to Psychology. The project explores how structured AI use influences research literacy, critical thinking, and perceptions of learning while emphasizing evidence-based and ethical AI integration in higher education.

- AI-Augmented Form Assembly from a Secure Item Bank: A Faculty Pilot**
Presenter(s): Fen Fan, National Board of Medical Examiners; Jerusha Henderek, NBME; Amanda Clauser, National Board of Medical Examiners
 This mixed-methods pilot examines how medical-school faculty leverage AI assistance to assemble forms from a secure item bank for a formative assessment program. Four faculty complete an AI-augmented build; pre- and post-task surveys, a post-task focus group, and (contingent) interaction logs capture perceived efficiency, form quality, and judgment boundaries.
- Everyday AI Use, Acculturation, and Collaborative Learning in Multicultural Students**
Presenter(s): Rocío López Caicedo, North Carolina State University; Andry Bustamante Barreto, Korea National University of Education
 This study investigates AI as cultural mediator by examining how everyday generative AI use relates to acculturation orientations and collaborative learning behaviors in multicultural university students. Using adapted acculturation, collaborative learning scales and newly developed AI exposure measures, we assess implications for equitable, measurement-informed AI integration in higher education.
- Evaluating Construct Alignment in AI-Generated Science Motivation Items from TIMSS Assessment Frameworks**
Presenter(s): Kazi Md Rukan Uddin, University of Houston; Idris Yusuf, University of Houston; Md. Safayet Alam, National Curriculum and Textbook Board, Bangladesh; Muhammad Yahya, Directorate of Secondary and Higher Education, Bangladesh
 Large language models are increasingly used to support assessment item development, yet limited evidence exists regarding construct preservation. Using TIMSS 2023 science motivation frameworks, 45 items generated by ChatGPT, Gemini, and Claude were evaluated. Preliminary findings indicate strong construct alignment, although redundancy and construct drift remain concerns.
- MIST: Misconception Identification and Sycophancy Tracking**
Presenter(s): Lionel Meng, University of Wisconsin - Madison; Lingbo Tong, University of Wisconsin-Madison; Icy Zhang, University of Wisconsin - Madison
 This study presents a framework for evaluating sycophantic behavior in AI tutors for statistics instruction. Investigation of pedagogical efficacy, sycophantic tendencies, and sensitivity of these factors to prompt strategies across Large Language Models are studied in simulated student-tutor conversations. Relevance to promoting trustworthy model behavior and effective learning is discussed.
- Building Teacher Capacity to Generate and Evaluate Math Items with Small LLMs**
Presenter(s): Minseok Kim, University of Pennsylvania
 Small language models can run offline in a classroom, but they make mathematical errors no teacher can afford to miss. This work-in-progress proposes a measurement-grounded professional development model in which upper-elementary teachers use local LLMs to generate and critically evaluate items; an illustrative pilot demonstrates feasibility, teacher-impact study planned.
- Evaluating Neural Network-Based IRT in Multistage Adaptive Testing with Small Item Banks**
Presenter(s): Seong Eun Hong, Center for Applied Linguistics
 This study investigates the performance of neural network-based IRT estimation in multistage adaptive language assessment with small item banks and short tests. Preliminary results indicate accurate item recovery despite limited training information. Accuracy improved with sample size and training iterations with diminishing returns, supporting efficient calibration with low computational cost

- Right Score, Wrong Reason? Rater Rationales and Construct Validity in LLM Scoring**
Presenter(s): Jia Quan, University of Kansas; Rebecca Swinburne Romine, University of Kansas; Kandace Fleming, University of Kansas; Russell Swinburne Romine, University of Kansas; Dale Cope, University of Kansas; Jonathan Schuster, ATLAS - University of Kansas
 Large language models can score essays with high agreement to human raters, but agreement alone does not tell us whether they are attending to the same construct-relevant features. Using scores and scoring rationales from the dataset, this work-in-progress study investigates potential “right score, wrong reason” cases in LLM essay scoring.
- Validating Creative Writing Scores: Do TTCT Verbal Subscores Predict LLM-Based Title-Generation Ratings?**
Presenter(s): Haeju Lee, University of North Carolina Greensboro; Sungyeun Kim, Incheon National University; Jinmin Chung, Univ. of Iowa
 This study examines whether TTCT Verbal subscores predict LLM-scored creative title-generation performance among students. AI scores will first be validated against human ratings via agreement and reliability indices. Regression and commonality analysis will then evaluate concurrent criterion-related validity, distinguishing task-general divergent-thinking signal from shared linguistic-method variance and AI-scoring artifacts.
- Prompt Engineering for LLM-Generated Likert-Scale Survey Responses**
Presenter(s): Nicole Bonge, American Board of Internal Medicine; Jihong Zhang, University of Arkansas at Fayetteville
 This study explores varied interview-informed prompt designs for LLM-generated survey responses across 24 prompt configurations and verification tasks (self-check, persona-check). Using the Teacher Stress Inventory as a case study, we find that prompt ordering impacted data quality more than verification tasks, which failed to bridge psychometric gaps.
- LLM-Based Metadiscourse Analysis of Demographic Variations in ELL Hedge Use**
Presenter(s): Tianyu Xu, Teachers College, Columbia University; Xinyu Xu
 Using an LLM-assisted pipeline, this study examines sociodemographic variation in English Language Learners’ hedge use across 6,500 essays. Regression analyses are expected to show that gender, race, and SES predict hedge density and category-specific variation, which subsequently predict writing scores, providing valid evidence for informing fair, instructionally meaningful writing assessment.
- Using Generative AI to Support Data-Drive Decision Making for Struggling Writers**
Presenter(s): Anya Evmenova, George Mason University
 This presentation examines the potential of generative AI to support providing writing feedback and instructional decision-making. Student essays were analyzed by teachers as well as by GenAI using different prompts. Anticipated findings include differences across AI conditions, partial alignment with teacher feedback, and implications for integrating AI into writing instruction.

FULL SCHEDULE

WEDNESDAY, OCTOBER 7

Keynote: Alina von Davier

Keynote

8:00 AM – 9:00 AM

Conference Center: Grand Ballroom 1

“The Computational Assessment Stack: Computational Psychometrics for AI-Native Measurement”

Artificial intelligence is transforming educational measurement, but its impact is often described one task at a time, say, an item generator or an automated scorer. This keynote argues that the deeper opportunity is architectural: computational psychometrics provides the scientific foundation for AI-native assessment systems in which every stage of the measurement lifecycle is represented computationally and designed to work together.

I propose the Computational Assessment Stack, a layered architecture for such systems. At its base is a computational representation of the construct and content domain: knowledge graphs, skill decompositions, and structured task models that make constructs machine-readable and comparable. Computational content generation, in which LLMs, agentic workflows, and validated scripts produce and review items, needs to be built on the construct representation. Then, computational psychometrics for delivery and estimation handles test assembly and adaptive administration as optimization problems over the same representations, and scoring as statistical and AI-based inference over rich response evidence. An orchestration layer or agent coordinates the stack and collaborates with the human experts, suggesting when to generate, calibrate, refresh, and retire content, and closing the loop between human-led and AI-led processes.

Drawing on examples from a high-stakes AI-native assessment, the Duolingo English Test, I examine what each layer of the stack can reliably accomplish today, where established psychometric methods remain competitive, and where open problems persist. The stack preserves the assessment pillars of validity, reliability, fairness, and transparency by design, guided by a single principle: every measurement decision should be represented computationally before it is automated. The result is a roadmap for building intelligent, auditable, and continuously improving measurement systems for the next generation (of learning and) assessment.

Classroom Implementation & Teacher Support

Individual Paper Session

9:15 AM – 10:45 AM

Conference Center: Smithfield/Birmingham

- **Evaluating Educators’ Instructional Skills on Making Content Explicit in Instruction in a**
Presenter(s): Shreyashi Halder, ETS; Jamie Mikeska, ETS; Heather Jorgenson, George Washington University
This paper examines mathematics educators’ instructional skills as they practice teaching strategies in a generative AI (GenAI) teaching simulation. Study findings indicated that educators were somewhat able to elicit student’s conceptual understanding and misunderstanding. While educators helped to address the student’s misconceptions, there remains a need for improvement.
- **An LLM-Enhanced Score Reporting Assistant for Teachers: System Design and Usability Evaluation**
Presenter(s): Shan Zhang, University of Florida; Caitlin Tenison, ETS; Diego Zapata-Rivera, ETS; Reginald Gooch, ETS; Maya Israel, University of Florida
This study introduces an LLM-powered Smart Report Assistant grounded in audience analysis and assessment design principles. Using retrieval-augmented generation, the system helps teachers interpret assessment data through personalized and conversational reporting. Results from a usability study indicate promise for supporting score interpretation and data-informed instructional decision-making.
- **A Specialized Multimodal Transformer for Classroom Discourse Classification**
Presenter(s): Chongyu He, University of Virginia; Jonathan Foster, University at Albany; Peter Youngs, University of Virginia; Scott Acton, University of Virginia
Multimodal classroom discourse classification becomes challenging in scenarios in which audio and video may

FULL SCHEDULE

WEDNESDAY, OCTOBER 7

be missing or noisy. We demonstrate a model that treats classroom transcript as the primary modality and leverages audio and video as incremental information sources. A diagnostic evaluation protocol for multimodal robustness will also be shared.

- **Designing for Understanding: A Learning-Science-Informed AI Application**

Presenter(s): Salih Mansur, Adelphi University

This innovation demonstration introduces a learning-science-informed AI application that transforms lengthy, fragmented chatbot responses into concise, structured, source-aware guidance. Drawing on cognitive load theory, multimedia learning, scaffolding, and human-centered design, it helps users identify essential information, improve comprehension, and move more effectively from information to understanding.

Interweaving Assessment: A Field-Generated R&D Agenda for AI and Measurement

Panel Session

9:15 AM – 10:45 AM

Conference Center: Commonwealth 2

This panel presents findings from Interweaving Assessment: AI, Assessment, and Accountability in Responsive Education, a three-convening initiative funded by the Gates Foundation and facilitated by Menlo Education Research. Drawing on structured input from over 150 researchers, practitioners, policymakers, and technology developers, the project produced an integrated R&D agenda organized around three clusters: innovating measurement, tools to support educational leaders, and datasets as a public good. Three cross-cutting themes, codesign, data sovereignty, and human-centered design, run across all priorities. While the full agenda addresses priorities across policy, practice, and measurement science, several priorities speak directly to the technical work of this community: validity frameworks for personalized and adaptive assessment, fairness audits for intersectional identities, multimodal construct definition, and the governance of openly licensed benchmark datasets. Panelists will present the agenda's substance, describe where measurement scientists are among the necessary partners in realizing it, and announce a proof-of-concept grant program to seed the highest-priority work. The session closes with structured audience engagement in which participants can provide input about priorities to those developing the grant program.

Moderator: Britte Cheng (Menlo Education Research)

- **Interweaving Assessment: A Field-Generated R&D Agenda for AI and Measurement**

Presenter(s): Britte Cheng, Menlo Education Research; Susan Lyons, National Council on Measurement in Education; Temple Lovelace, Assessment for Good, AERDF; Lauren Deters, Khan Academy; Lauren White, WestEd; John Whitmer, Learning Data Insights

LLMs for Mathematics Understanding and Scoring

Individual Paper Session

9:15 AM – 10:45 AM

Conference Center: Brigade

- **LLM-Generated Distractor Rationales for Explanations of the Misconceptions in Mathematics Assessment Items**

Presenter(s): Jiawei Xiong, Curriculum Associates; Erin Banjanovic; Catherine Feuerstein, Curriculum Associates

This study develops a proof-of-concept framework using large language models to generate distractor rationales, to explain the misconceptions reflected in incorrect answer choices for mathematics items. The approach integrates item content, metadata based on artifacts from principled assessment design (PAD), and expert validation to enhance the diagnostic value of assessments.

- **Automated Approaches for Scoring Math Misunderstandings in Student Self-Explanations**

Presenter(s): Scott Crossley, Vanderbilt University; Bethany Rittle-Johnson, Vanderbilt University; Rebecca Adler, Vanderbilt University; L Burleigh, The Learning Agency; Kelley Durkin, Vanderbilt University; Jules King, The Learning Agency; Meg Benner, The Learning Agency

This study reports on an open data science competition based on a benchmark dataset of mathematics misunderstandings, comprising over 52,000 mathematics explanations written to justify answer choices from 15 multiple-choice questions that were labeled by expert human annotators. Competitors were tasked with correctly classifying the explanations and any misunderstandings.

- **Measurement-Informed Difficulty Priors for Competition Math Items**

Presenter(s): Amirreza Mehrabi, Purdue University; Breejha Quezada, Purdue University; Jason Morpew, Purdue University

This WIP tests whether embeddings, LLM-extracted solution features, and LLTM-inspired concept indicators can support early difficulty priors for competition mathematics. Initial results show consistent LLM-feature gains and an emerging role for concept indicators as measurement-grounded interpretation.

- **Generalizability of Automated Scoring of Mathematical Proofs with AI Examinees and Raters**

Presenter(s): Lixin Wu, University of Illinois at Urbana-Champaign; Zhimeng Luo, University of Pittsburgh; Jinming Zhang, University of Illinois at Urbana-Champaign

As AI writes and scores open-ended work, scoring dependability is unclear. Using generalizability theory, we decompose score variance in LLM scoring of AI-written mathematical proofs across item, judge, rubric, and replication facets. Preliminary results show examinee rankings robust to judge and rubric choice but limited by six items.

- **Making Mathematics Problems More Approachable Through AI Context Personalization**

Presenter(s): Daiki Hojo, Benesse Educational Research and Development Institute; Yoshio Goto, Benesse Educational Research and Development Institute; Hiroyuki Maruyama, Benesse Educational Research and Development Institute; Yuki Shirasuna, Benesse Educational Research and Development Institute; Akihiro Sato, Benesse Educational Research and Development Institute; Kentaro Kato, Benesse Educational Research and Development Institute

This demonstration presents an AI prototype that transforms mathematics items into learner's interest contexts while preserving their underlying mathematical structure to support approachability and engagement. We discuss early technical verification, teacher review, and iterative refinement of criteria for mathematically coherent personalization.

LLMs for Multilingual Content Scoring

Individual Paper Session

9:15 AM – 10:45 AM

Conference Center: Duquense

- **Evaluating Multilingual Language Models for Scoring Mixed English—Spanish Assessment Items**

Presenter(s): Joshua Southerland, Pearson; Martha Bellows, Pearson; Bradley Behan, Pearson; Susan Lottridge, Pearson

We evaluate whether fine-tuning pretrained language models supports automated scoring in a bilingual setting, comparing MrBERT-es and XLM-RoBERTa Base across 31 items. Both models achieve reliable agreement for most items, enabling unified scoring without language detection, with MrBERT-es providing better performance overall.

- **Comprehensive Pipeline for Multilingual GenAI Scoring**

Presenter(s): Ji Yoon Jung, Boston College; Ummugul Bezirhan, Boston College; Matthias von Davier, Boston College

This study presents a comprehensive operational pipeline for multilingual GenAI scoring that integrates secure data preparation, prompt generation, automated scoring, post-processing, and human-in-the-loop review. Results demonstrate the pipeline's strong adaptability across item formats, assessment languages, and assessment cycles, offering a viable solution to the persistent challenges of multilingual scoring.

- **Lost Without Translation? Multilingual Sentence Embeddings for Linguistic-integrated Reliability Auditing**

Presenter(s): Ummugul Bezirhan, Boston College; Ji Yoon Jung, Boston College; Matthias von Davier, Boston College

Multilingual assessment systems commonly rely on translation for scoring and quality-control processes. We evaluate whether multilingual sentence embeddings can replace translation for reliability auditing across 11 PIRLS constructed-response items and three embedding models. Reliability estimates remained highly similar while untranslated responses were recovered without measurable impact on reliability.

- **Designing and Evaluating a Measurement-Oriented RAG Framework for ELL Writing Assessment**

Presenter(s): Daeryong Seo

This study designs and evaluates a tool-augmented, measurement-oriented RAG workflow for ELL writing assessment. The design integrates question routing, measurement tools, controlled retrieval, and evidence hierarchy. Across three repeated evaluation rounds, the system showed stable tool use, evidence grounding, rubric interpretation, and measurement-safe reasoning related to validity, reliability, and fairness.

- **Mitigating Bias in Automated Essay Scoring for ESL Learners via Contrastive Learning**

Presenter(s): Kevin Fan, Johns Creek High School; Georgia Institute of Technology; Eric Yun, Georgia State University

Automated Essay Scoring systems disproportionately penalize high-proficiency English as a Second Language (ESL) learners. We propose Contrastive Learning with Matched Essay Pairs (CL-MEP), a bi-directional alignment strategy. CL-MEP reduces this scoring bias by 39.9% while improving overall accuracy, successfully disentangling valid syntactic complexity from surface-level grammatical errors.

FULL SCHEDULE

WEDNESDAY, OCTOBER 7

Leveraging LLMs for Process Data Analysis and Scoring

Individual Paper Session

9:15 AM – 10:45 AM

Conference Center: King's Garden 5

- **Finding Evidence of Communication Behaviors using LLMs**

Presenter(s): Tazin Afrin, NBME

This study compares two automated scoring approaches to identify communication behaviors in physician responses to patient questions: prompt-based scoring with LLM and supervised transformer-based models. Our results show that although transformer-based models provide more consistent performance, competitive LLM performance is promising in reducing the need for annotated training data.

- **Measuring Writing Proficiency Through Keystroke Process Data: A Cross-School Generalizability Study**

Presenter(s): Damilola Babalola; Naga Buddarapu, North Carolina State University; Caleb Scott, North Carolina State University; Piotr Mitros, Educational Testing Service; Paul Deane, ETS; Collin Lynch, North Carolina State University

Using a year of authentic Grade 7–8 writing from two U.S. middle schools (8,576 documents, 423 students), we test whether keystroke features correlate with English Language Arts proficiency, form interpretable latent dimensions, and vary by gender. Fluency and engagement features reliably index proficiency by Grade 8, though generalizability varies.

- **Reliability Estimation Methods for Bayesian Knowledge Tracing Mastery Classifications**

Presenter(s): Samuel Ihlenfeldt, Carnegie Learning; Stephen Fancsali, Carnegie Learning, Inc.

We evaluate the psychometric reliability of Bayesian Knowledge Tracing for skill mastery classifications. Using large-scale intelligent tutoring system data, we compare two simulation-based approaches for estimating classification consistency as reliability. Results support both approaches, show reliability increases with response-sequence length, and highlight the necessity of adjusting for chance agreement.

- **pairadigm: Measurement of Scalar Constructs Using Large Language Models**

Presenter(s): Michael Chrzan, Center for Educational Data Science and Innovation, University of Maryland

pairadigm is a free, open-source Python package providing an end-to-end pipeline for LLM-assisted construct annotation of text. It integrates Concept-Guided Chain-of-Thought prompting, Bradley-Terry scoring, and the Alternative Annotator Test to help researchers generate continuous construct scores from text while rigorously validating when LLM annotations can replace human judgment.

New Frontiers in Automated Essay Scoring

Individual Paper Session

9:15 AM – 10:45 AM

Conference Center: Heinz

- **Score Comparability Limits with Common AES Validation Measures**

Presenter(s): Tim Moses, Buros Center for Testing; Seongeun Kim, National Commission on Certification of Physician Assistants; Nick Trout, Michigan State University

Automated Essay Scoring (AES) systems are often validated using measures of agreement in AES and human scores. This paper shows that differences in scoring standards can occur even when AES scoring meets established thresholds in widely used agreement measures. Demonstrations are provided for winners of recent national AES competitions.

- **Reference-Conditioned Evaluation of Automated Essay Scoring**

Presenter(s): Yong He, Measurement Incorporated; Corey Palermo, Measurement Incorporated

Automated essay scoring is typically evaluated using human–machine differences, but these cannot distinguish

machine from human sources of disagreement. This study integrates embedded reference-response evidence to support attribution and operational decision making. Simulations identify conditions for reliable operational implementation.

- **An Evaluation of Agreement and Uncertainty in LLM-Based Automated Essay Scoring**

Presenter(s): Yiting Yao, Florida State University; Yanyun Yang, Florida State University; Hailey Kuang, Florida State University

This study evaluates score agreement and uncertainty estimation in large language model (LLM)-based automated essay scoring. Results showed that LLMs tended to be overconfident and highly consistent in their assigned scores, and majority voting did not yield stronger agreement with human scores than deterministic scoring.

- **LLM-Human Agreement in AES: A Meta-Analysis of Studies Using TOEFL11**

Presenter(s): Hongli Li, Georgia State University; Che-Han Chen, Georgia State University; Chiho Young-Johnson, Georgia State University; So-Young Lim, Georgia State University; Yali Feng; Kevin Fan, Johns Creek High School; Georgia Institute of Technology

This meta-analysis synthesized studies examining agreement between large language models (LLMs) and human raters in automated essay scoring using the TOEFL11 corpus. A three-level meta-analysis revealed moderate LLM-human agreement and substantial heterogeneity both between and within studies. Newer-generation LLMs, fine-tuning, and adaptation procedures were associated with higher agreement.

- **Bridging Human and Automated Scoring in Large-Scale Writing Assessment**

Presenter(s): Ru Lu, Educational Testing Service; Venessa Manna; Naim Alper, ETS; Klaus Zechner, ETS; Cady Gansen, ETS

Using operational data, this study evaluates NLP-, LLM-, and hybrid, including human scoring for written responses in a large-scale language assessment. Results show that automated models improve scoring consistency but produce larger differences in scale scores and classifications. Hybrid human-machine models better align with human scoring while also improving reliability.

Short-Answer & Constructed-Response Scoring

Individual Paper Session

9:15 AM – 10:45 AM

Conference Center: Ft. Pitt

- **Fine-tuning Large Language Models for Automated Scoring: Classification, Regression, vs. Ordinal Regression**

Presenter(s): Jiyun Zu, ETS; Akshay Badola, ETS

Fine-tuning large language models for automated scoring is often formulated as either a regression or a classification task. However, scores assigned by human raters are on an ordinal scale. We summarize different deep-learning ordinal regression methods and compare their performances with those from regression and classification using a real dataset.

- **Examining LLM-Surprisal as an Indicator of Naturalness for Japanese Automated Essay Scoring**

Presenter(s): Akari Osumi, Purdue University; Jingying Hu, Purdue University; Yan Cong, Purdue University; Atsushi Fukada, Purdue University

This study investigates Large Language Model (LLM) surprisal as an indicator of global linguistic naturalness in L2 Japanese automatic essay scoring. Results demonstrate that surprisal effectively distinguishes learner proficiency levels. Combining surprisal with classic feature-based indices maximizes proficiency classification accuracy, validating surprisal as a robust metric for automated writing evaluation.

- **Autoscoring Anticlimax: A Meta-analytic Understanding of AI's Short-answer Shortcomings and Wording Weaknesses**

Presenter(s): Michael Hardy, Stanford University

We meta-analyze 890 culminating results across a systematic review of LLM short-answer scoring studies. We estimate the factors that contribute to LLM performance, quantifying the disconnect between human and LLM difficulties in SAS and provide recommendations for developers working on SAS for schoolchildren.

- **Sample-Efficient Fine-Tuning for Operational Short-Answer Scoring: A Measurement-Anchored Evaluation**

Presenter(s): Joshua McGrane, University of Melbourne; David Torres Irribarra, Pontificia Universidad Catolica de Chile

We fine-tune three LLMs (GPT-4o-mini, gpt-oss-20b, gpt-oss-120b) to score ASAP-SAS short answers across six exemplar-budget sizes, evaluating agreement with both QWK and PRMSE against operational standards. Fifty exemplars per score class clears the Williamson–Xi–Breyer panel on most items, and the three models are statistically indistinguishable; a small open-weight model suffices.

- **Information-Theoretic Dimension Reduction for LLM Embeddings in Automated Scoring**

Presenter(s): Xiang Liu, ETS; Michael Fauss, ETS

We propose an information-theoretic framework for dimension reduction of high-dimensional LLM embeddings in automated scoring. By maximizing mutual information between projections and scores using a KSG estimator and simulated annealing, the method identifies low-dimensional representations that preserve construct-relevant information while improving interpretability and robustness.

Reexamining Interactivity: How AI Transforms and Scales Modern Language Assessment

Sponsor Showcase

9:15 AM – 10:45 AM

Conference Center: Commonwealth 1

Modern technology allows test developers to potentially improve the construct representation of assessment tasks. For instance, large language models (LLM) process natural language and produce dynamic output, allowing LLMs to be used to simulate conversations. Test developers must ensure that the resulting tasks adequately measure the intended construct. The purpose of this talk is to explain how AI-enabled assessment items are developed and validated for use on the Duolingo English Test (DET) using the Interactive Speaking task as an example.

The Interactive Speaking task balances standardization and adaptivity. Subsequent questions depend on test-taker responses to earlier questions, facilitating a degree of adaptivity. Specifically, follow-up questions are selected based on LLM-enhanced evaluation of how well responses align with question-specific rubrics. The responses are then graded on six subconstructs using NLP-derived features and supervised learning models and scored using a six-dimensional continuous response IRT model. This structure supports standardized administration and consistent scoring while eliciting more conversational language than existing monologic tasks. Moreover, the IRT-based scores were more reliable than non-IRT grades for the same responses, and the interactive task produced more reliable scores than the monologic task with greater consistency across the language proficiency groups.

We first describe the motivation for the Interactive Speaking task on the DET. We then review the steps required to determine how AI could be used to produce item content, process test-taker responses, and select follow-up questions. We next detail the process for grading and scoring responses to these items. Specifically, we present software developed to enable IRT-based calibration and scoring of complex and innovative item types using features extracted from item content. Finally, we explain how this development process could be applied to other interactive tasks to support more representative response processes while maintaining reliable and construct-relevant scores. Moderator: Steven Nydick (Duolingo) and Yena Park (Duolingo)

AI Literacy

Individual Paper Session

11:00 AM – 12:30 PM

Conference Center: Ft. Pitt

- **Using AI to assist with the development of an AI literacy assessment**

Presenter(s): Diego Figueiras, ETS; Guangming Ling, ETS

We propose using AIGENIE to develop an AI literacy assessment following the Adapt AI Reflect framework. The AIGENIE-generated assessment will be evaluated on psychometric properties such as content coverage, internal structure, reliability, item statistics, and test information functions in comparison with Adapt AI Reflect.

- **Measuring What Matters: Construct Coverage in AI Literacy Assessment**

Presenter(s): Jodi Davenport, WestEd; Yvonne Kao, WestEd; Sarah Quesen, WestEd

AI literacy assessments have proliferated, but what do they measure? We analyze construct coverage across 13 assessments and 14 frameworks along five flourishing-derived dimensions. Preliminary results show systematic underrepresentation, with human connection absent throughout. We frame these gaps as validity threats and propose an agenda for next-generation measurement.

- **Measuring Science Teachers' Generative AI Literacy for Classroom Assessment: Performance Assessment**

Presenter(s): Ruiping Huang, University of Illinois Chicago; Yue Yin, University of Illinois at Chicago

We assessed science teachers' generative AI (GenAI) literacy for classroom assessment using performance-based tasks and compared the results with self-reported measures of GenAI use frequency and confidence. Preliminary findings indicate a moderate correlation with use frequency but only a weak correlation with confidence, highlighting the value of performance assessments.

- **Developing the AI Output Engagement Scale**

Presenter(s): Youlim Lee, University of Massachusetts Lowell; Jill Hendrickson, University of Massachusetts Lowell

This study presents preliminary validity evidence for the AI Output Engagement Scale, a 30-item instrument measuring how graduate students notice, revise, and review AI-generated content. Exploratory and confirmatory analyses provide initial support for a three-factor structure, while item diagnostics identify specific revisions needed before broader use.

- **Measuring the User, Not the Model: Critical AI Literacy**

Presenter(s): Derek Briggs, University of Colorado - Boulder

The AI community speaks of “measuring” large language models, but measurement presupposes variability that LLM development eliminates. The model is an instrument, not a measurable object; measurement relocates to the human—critical AI literacy. A simulation shows this construct is most measurable where verification is expensive and AI helps least.

Demonstrating Validity Evidence for LLM-Based Assessment Tools**Individual Paper Session****11:00 AM – 12:30 PM****Conference Center: Duquense**

- **Measuring AI-Generated Feedback: Validity Evidence for AI-Integrated Pedagogy**

Presenter(s): Yetunde Fawehinmi, Texas A&M University

This conceptual paper proposes a validity evidence framework for evaluating AI-generated formative feedback in higher education classrooms. It identifies evidence domains: construct representation, feedback accuracy and alignment, learner response process, fairness and access, and pedagogical uptake. The framework supports defensible claims about assessment quality, learning support, and instructional decision-making practices.

- **Implementing Generative AI Assessment Guidelines in a Mexican University**

Presenter(s): Melchor Sánchez-Mendiola, National Autonomous University of Mexico (UNAM); Enrique Buzo-Casanova, National Autonomous University of Mexico (UNAM); Elibidú Ortega-Sánchez, National Autonomous University of Mexico (UNAM); Manuel García-Minjares, National Autonomous University of Mexico (UNAM); Adriana Cruz Juárez, National Autonomous University of Mexico (UNAM)

This case study examines the design and launch of guidelines for generative AI use in assessment across high school, undergraduate, and graduate education in a large public university in Mexico. We analyze a collaborative, level-specific policy process and derive implementation lessons for validity, equity, academic integrity, transparency, and faculty development.

- **A Validity Argument Framework for Automated Item Generation Systems**

Presenter(s): Euigyum Kim, Sogang University; Hyo Jeong Shin, Sogang University; Alina von Davier, Duolingo

This study introduces an evidence-centered, argument-based framework for validating AI-generated items in educational assessments. By synthesizing a series of claims, assumptions, and evidence within a stepwise and integrative validity argument, we show that AI-generated items can establish robust validity evidence comparable to human-authored items through distinct underlying mechanisms.

- **What an Expert Evaluation of an LLM Tutor Reveals and Conceals**

Presenter(s): Farzan KarimiMalekabadi, Edmentum

We report a completed expert evaluation of a deployed LLM tutor for grades 6–12: 260 sessions scored by 12 subject-matter experts. Aggregate ratings were high (86% judged comfortable for students), but coverage skew, single-rater scoring, and composite averaging limit what the favorable numbers establish about safety and reliability.

- **Applying Evidence-Centered Design to Automated Evals of AI-Powered Assessment Systems**

Presenter(s): Kristen Dicerbo, Khan Academy; Britte Cheng, Menlo Education Research; John Whitmer, Learning Data Insights

This paper demonstrates Evidence-Centered Design as a principled approach to design automated evaluations of AI-powered assessment outputs. We demonstrate this application through Khan Academy's "Explain Your Thinking" activity, showing how ECD's layered models can be translated into interpretable systems to demonstrate validity in AI systems.

Educational Measurement as an AI-Native Profession

Panel Session

11:00 AM – 12:30 PM

Conference Center: Commonwealth 2

Artificial intelligence is not only changing assessment products; it is changing the professional work of educational measurement itself. This panel examines how an AI-native future may transform day-to-day practice, professional norms, and epistemic responsibility in the field. Bringing together perspectives from education measurement experts from academia, assessment industry, professional leadership and independent consulting, the panel will address several aspects in which AI intersects with current work including: scholarly publishing and peer review, assessment development and operations, graduate training, integrated psychometric and data-science workflows, documentation and reporting. The session emphasizes concrete use cases rather than general speculation: how AI is already changing technical work, where it can responsibly accelerate, deepen, critique, or audit professional judgment, and where human accountability sits currently in this matrix. The discussion will focus on emerging standards for validity, fairness, transparency, reproducibility, attribution, quality control, and public trust as AI becomes embedded in measurement practice. The goal is to move from tool-centered conversations toward a profession-level account of how educational measurement should adapt while preserving rigor and human responsibility.

Moderator: Andrew Ho (Harvard University)

- **Educational Measurement as an AI-Native Profession**

Presenter(s): Damian Betebenner, Center for Assessment; Frank Rijmen, Cambium Assessment; Derek Briggs, University of Colorado - Boulder; Mohammed Abulela, MetaMetrics, Inc. and University of Minnesota

Artificial intelligence is not only changing assessment products; it is changing the professional work of educational measurement itself. This panel examines how an AI-native future may transform day-to-day practice, professional norms, and epistemic responsibility in the field. Bringing together perspectives from education measurement experts from academia, assessment industry, professional leadership and independent consulting, the panel will address several aspects in which AI intersects with current work including: scholarly publishing and peer review, assessment development and operations, graduate training, integrated psychometric and data-science workflows, documentation and reporting. The session emphasizes concrete use cases rather than general speculation: how AI is already changing technical work, where it can responsibly accelerate, deepen, critique, or audit professional judgment, and where human accountability sits currently in this matrix. The discussion will focus on emerging standards for validity, fairness, transparency, reproducibility, attribution, quality control, and public trust as AI becomes embedded in measurement practice. The goal is to move from tool-centered conversations toward a profession-level account of how educational measurement should adapt while preserving rigor and human responsibility.

Evidence-Centered Design Anchored AI: Strengthening Measurement Science in STEM

Coordinated Paper Session

11:00 AM – 12:30 PM

Conference Center: Heinz

Artificial intelligence (AI) offers powerful capabilities to accelerate assessment development, automate item generation and scoring, and provide personalized feedback to support student learning. However, its use also introduces challenges around validity, fairness, and transparency. Evidence-Centered Design (ECD) provides a principled framework for integrating AI throughout the assessment cycle.

This session showcases how ECD can serve as an organizing architecture for AI-supported assessment across various STEM contexts. It includes one conceptual presentation with four empirical studies that examine AI applications across STEM contexts, including multimodal scoring of constructed responses, equitable assessment of culturally and linguistically diverse students, and the use of a to provide personalized feedback.

The five presentations highlight ECD-guided innovations such as AI-generated mathematical modeling tasks, formative assessments and credentialing in semiconductor engineering, large language model scoring of open-ended software engineering responses, and AI scoring of mixed-medium engineering design tasks.

By anchoring AI within a rigorous psychometric approach, ECD promotes valid, fair, and transparent assessment practices-directly addressing the AIME-Con theme of Measurement Science in AI-Integrated Assessment and Pedagogy.

Discussant: Andre Rupp (Center for Assessment)

Chair: Maria Oliveri (Purdue University)

- **Evidence-Centered Design as Anchor for Principled AI Integration in STEM Assessment**
Presenter(s): Maria Oliveri, Purdue University; Eric Tucker, The Study Group
- **AI-Assisted Evidence-Centered Design to Build a Blueprint for a Semiconductor Credential**
Presenter(s): Emily Proctor, College of Engineering, Department of Engineering Education, Purdue University; Maria Oliveri, Purdue University; Sarah Hammami, University of Nebraska-Lincoln; David Foster
- **Evidence-Centered Design for AI-Driven Automated Item Generation in Applied Mathematics**
Presenter(s): Edith Graf, Proof-of-Concept Learning LLC; Maria Oliveri, Purdue University; Sarah Hammami, University of Nebraska-Lincoln
- **Using Large Language Models for Scoring Open-Ended Responses in Software Engineering**
Presenter(s): Alex Rudniy, Drew University
- **AI Scoring of Multimodal Student Work in Multi-Staged Project-Based Learning Assignments**
Presenter(s): Eileen Koh, Purdue University; Melissa Dyehouse, Purdue University; Alex Rudniy, Drew University; Maria Oliveri, Purdue University; Sarah Hammami, University of Nebraska-Lincoln

FULL SCHEDULE

WEDNESDAY, OCTOBER 7

Human and AI Partnerships for Learning and Assessment

Individual Paper Session

11:00 AM – 12:30 PM

Conference Center: King's Garden 4

- **Teacher Implementation of AI-Supported Assessment for Scientific Argumentation and Responsible AI Use**

Presenter(s): Field Watts, ETS; Lei Liu, ETS; Amy Adair, ETS; Yi Song, ETS Research Institute

This work-in-progress presents a case study of a middle school science teacher implementing an AI-supported formative assessment of scientific argumentation. Using the Responsible and Ethical Principles framework, we examine how teacher mediation shaped students' engagement with AI and the interpretation of assessment evidence, informing design and validity considerations.

- **Principled Approaches to Building AI Partners for Assessing and Supporting Small-Group Collaboration**

Presenter(s): Peter Foltz, University of Colorado - Boulder

We present a principled, evidence-centered design approaches for building AI partners that assess and support classroom collaboration. Using two implementations, the Community Builder (CoBi) and the Jigsaw Interactive Agent (JIA), we illustrate how decisions about evidence sources, granularity, timing, and feedback target trace back to a shared measurement framework.

- **Measuring Collaborative Reasoning with LLMs**

Presenter(s): Dawei Xie, Northwestern University; Tochukwu Eze, Northwestern University; Marcelo Worsley, Northwestern University

We present CRS, a discourse framework representing multi-party reasoning as contributions, relations, and derived structure. Evaluating five LLMs on student discussions, we find reasoning is classifiable but not reliably segmentable where identifying contributions is the bottleneck. Prompting improves labeling, not identification; participation is recoverable, yet fine-grained structure is hard.

- **An exploration of using an AI-based approach to shorten an assessment**

Presenter(s): Guangming Ling, ETS; Diego Figueiras, ETS

This study explores an AI-based approach to shorten a personality assessment and compares the content/construct representations, as well as a series of psychometric properties, of the AI-reduced assessment with those of a similarly shortened assessment by human experts.

Impacts of AI on Educators, Students, and Assessment

Individual Paper Session

11:00 AM – 12:30 PM

Conference Center: Smithfield/Birmingham

- **A Validity Threat Framework for Measuring Student Generative AI Use**

Presenter(s): Olukayode Apata, Texas A&M University; Yetunde Fawehinmi, Texas A&M University; Daniel Oyeniran, The University of Alabama; Glory Saidu, University of Minnesota; Segun Ajose, University of Kentucky; Naphtali Onalo, University of Ibadan; Oluwasegun Amoniyon, University of Pittsburgh

This conceptual paper presents a validity threat framework for measuring students' GenAI use in higher education. It identifies threats related to construct definition, response processes, self-report bias, policy context, fairness, and score interpretation, offering practical guidance for stronger measurement, assessment, policy, and pedagogy in AI-integrated education.

- **High School Students' Generative AI Over-reliance and Learning Goal Orientation**

Presenter(s): Helena Liu, West Lafayette Junior/Senior High School; Newnew Deng, West Lafayette Junior/Senior High School; Hongyi Zhai, North Oconee High School; Wenchao Ma, University of Minnesota

This study examines the relationship between high school students' reliance on GenAI and potential predictors:

student demographic characteristics, academic background, and learning goal orientation. The findings highlight the importance of guiding students to use GenAI appropriately, preventing over-reliance on AI, and fostering mastery goal orientations in the AI age.

- **Faculty perceptions of the impact of AI on the validity of assessment**

Presenter(s): David Miller, University of Florida; Tammie Cumming, City University of New York Baruch College
We examined faculty perceptions of construct change and assessment validity in the presence of generative AI. Using a cross-institutional survey (N = 438), faculty believed generative AI had strong impact on assessments, resulting in widespread construct change, only moderate confidence in assessments, and limited assessment redesign of assessment practices.

- **Trust Calibration in AI-Assisted Content Verification**

Presenter(s): Laine Bradshaw, Pearson
When AI-assisted verification tools generate potential concerns in assessment content, how do experts respond? This study examines SME responses to AI-generated verification concerns across operational assessment development workflows, providing early evidence about trust calibration, human-AI collaboration, and effective quality assurance processes in educational measurement.

Leveraging Large Language Models for More Personalized and Authentic Assessments **Coordinated Paper Session**

11:00 AM – 12:30 PM

Conference Center: Grand Ballroom 1

Artificial intelligence is increasingly embedded in how assessments are designed, administered, and interpreted, reshaping core psychometric principles and practices (Bulut et al., 2025; Sireci, Araneda, & Suárez-Álvarez, 2026; Suárez-Álvarez et al., 2026). Its psychometric value extends beyond cost savings, and while the evidence base is still emerging, AI may enable more personalized and authentic assessments. However, compared with automated item generation and scoring that often dominate large-scale assessment agendas, less attention has been given to assessments that feel realistic and engaging. Addressing this gap requires measurement science to engage with the technical foundations of AI in assessment and pedagogy, including natural language processing, retrieval architectures, agentic workflows, and immersive virtual reality (VR).

This symposium features four papers on how measurement science can evaluate, constrain, and improve AI-enabled systems to support more personalized and authentic assessments. The papers present a framework for evaluating prompt–output alignment, an empirical comparison of retrieval-augmented generation for culturally and workplace-relevant reading passages, a multi-agent method to validate and refine personalized math problems, and a topic-modeling systematic review mapping VR use in educational assessment and highlighting key gaps and future directions for standardized testing.

Discussant: Okan Bulut (University of Alberta)

Chair: Javier Suárez-Álvarez (University of Massachusetts, Amherst)

- **A Framework to Evaluate AI Workflows for Personalized Adult Skills Assessments**

Presenter(s): Anna Nasyrova, University of Massachusetts, Amherst

- **Comparing Retrieval Architectures for Enhancing Personalized Adult Reading Assessments**

Presenter(s): Haneul Lee, University of Massachusetts Amherst

- **A Multi-Agent Approach to Validate and Refine LLM-Generated Personalized Math Problems**

Presenter(s): Fareya Ikram, University of Massachusetts Amherst

- **Integrating Virtual Reality in Educational Assessment: A Systematic Review with Topic Modeling**

Presenter(s): Qiwei He, Georgetown University

Multimodal and Visual Assessment

Individual Paper Session

11:00 AM – 12:30 PM

Conference Center: Brigade

- **Validating LLM-Assisted Scoring for Multimodal Geospatial Reasoning Tasks in STEM Assessment**

Presenter(s): Qingzhou Shi, Northwestern University; Lauren Smith, Northwestern University; Bob Kolvoord, James Madison University; David Kraemer, Dartmouth College; Adam Green, Georgetown University; David Uttal, Northwestern University

This study develops and evaluates a hybrid scoring pipeline for multimodal geospatial reasoning tasks. It combines rule-based analysis of students' geospatial model revisions with LLM-based scoring of written reasoning. Human-LLM comparisons across 174 responses provide preliminary validity evidence for scoring model accuracy, reasoning quality, and model consistency.

- **Does including images improve multimodal language models' accuracy on item discrimination estimation?**

Presenter(s): Jae Jun Jong; Miryeong Koo, University of Illinois at Urbana-Champaign

This study measures the effectiveness of including images to predict discrimination parameters of reading items. The result suggests providing images to language models is beneficial. Linear regression models based on both images and texts could explain data better. Random forests (Breiman, 2001) based on images and texts showed higher accuracy.

- **Image-Based Representation of Item Response Patterns for Test integrity**

Presenter(s): Akshay Badola, ETS; Mo Zhang, Educational Testing Service; Ikkyu Choi, ETS

This work re-frames anomalous test response detection in standardized testing as binary image classification by transforming test response logs into fixed-size matrix representations and treating them as grayscale images. Deep learning architectures, including ResNet and Vision Transformer, are evaluated. Experiments demonstrate that the approach is feasible and achieves high Recall.

- **Listening for Meaning: Evaluating an ASR-LLM Pipeline for Thematic Audio Analysis**

Presenter(s): Jada Willse, James Madison University; Brian Leventhal, James Madison University; Brian French, Washington State University

This proof-of-concept study evaluated whether an ASR-LLM pipeline can transcribe student audio and classify transcripts into themes relevant for culturally responsive assessment development. Background noise strongly affected transcription and downstream classification. Fine-tuned RoBERTa outperformed Copilot prompting, suggesting scalable promise with careful evaluation and human oversight.

Sponsor Showcase

11:00 AM – 12:30 PM

Conference Center: Commonwealth 1

Using LLMs in the Item Writing Process

Individual Paper Session

11:00 AM – 12:30 PM

Conference Center: King's Garden 5

- **Self-Revising Agents as Item Writers: Benchmarking Against Professional Item Writing**
Presenter(s): Steven Tang, eMetric LLC; Zhen Li; Richard Patz, University of California, Berkeley
This study compares a self-revising multi-agent AI framework to expert humans in developing a robust bank of items for a university calculus course. Different Large Language Models (LLM) and different prompt inputs are examined, including single versus double item exemplars. Both AI- and human-written items are evaluated on 10 criteria.
- **Production-Ready Automated Item Generation in Educational Assessment: Integration with Operational Workflows**
Presenter(s): Hotaka Maeda, Smarter Balanced; Kargi Chauhan, Smarter Balanced; Tharunya Chandrashekar, Smarter Balanced
We present a production-ready automated item generation pipeline, integrated into large-scale assessment program workflows. Our one-shot approach uses a sample item and established guidelines to prompt LLMs to generate items that follow all requirements. Our system automatically populates metadata, screens content quality, and uploads them to a live item bank.
- **Secure Use of AI Agents for Item Generation in K-12 - Large-scale Assessments in the Public Sector**
Presenter(s): John Workman, EQAO
This demonstration presents secure Copilot-based AI agents for item generation. Emphasizing enterprise data protection, psychometric rigour, and human-in-the-loop workflows, it shows how AI-generated items can be securely developed, reviewed, field-tested, and integrated into operational large-scale assessments.
- **A Comprehensive Evaluation of GenAI Based Items in a Medical Examination**
Presenter(s): Yanlin Jiang, NCCPA; Marcus Walker, National Commission on Certification of Physician Assistants; Andrew Dallas, NCCPA; Aquia Richburg, National Commission on Certification of Physician Assistants; Nikole Gregg, National Commission on Certification of Physician Assistants; Brittany Corrigan, National Commission on Certification of Physician Assistants
The study seeks to evaluate GenAI based items developed for medical assessments. The findings are expected to advance understanding of how LLMs can be effectively integrated into item generation and provide evidence to inform better practices regarding the development, validation, and use of GenAI based items in high-stakes assessments.
- **Off the assembly line: Applying manufacturing benchmarks to real-time assessment generation**
Presenter(s): Kjorte Harra, WGU Labs; Megan Imundo, WGU Labs; Lesley Reilly, WGU Labs; Cory Hammon, WGU Labs; Lauren Zito; Betheny Gross, WGU Labs; Catrina Nieser, WGU Labs
Demonstrating improvement in assessment design means validating the system generating each item just-in-time, and the test it assembles for each learner, rather than a fixed item pool. Borrowing from manufacturing quality control, we model composite item failure as a Poisson process over test length, demanding per-item quality beyond conventional thresholds

AI Competencies in Educational Measurement: Expanding the NCME Foundational Competencies Coordinated Paper Session

1:45 PM – 3:15 PM

Conference Center: Commonwealth 1

In 2024, the NCME Foundational Competencies in Educational Measurement (FCEM) framework largely relegated AI to “technical, statistical, and computational competencies.” As AI transforms measurement practice, the field is now addressing urgent professional and pedagogical questions: What should measurement professionals be able to do with AI, how should we teach it, and how do we know we have learned it?

Kadriye Ercikan will chair the symposium and update attendees on her NCME Presidential Task Force on AI-Integrated Assessment. Michael Russell, a member of the original FCEM Task Force, will argue that AI elevates the stakes of existing competencies rather than replacing them. Maggie Beiting-Parrish will present a framework linking existing FCEM competencies to AI and ML applications. Susan Davis-Becker will address the defensibility challenge facing credentialing organizations seeking to set AI competency standards when professional AI adoption is variable, evolving, and role-specific. And Matthew Johnson will posit that technical AI competencies are measurement competencies, using an example of a scoring task that asks learners to justify their agreement metric, run a fairness check, and detect bias. Andrew Ho, chair of the original FCEM Task Force, will discuss, moderate presenter cross-talk, and ensure presentations are sufficiently brief to allow plentiful audience engagement.

Discussant: Andrew Ho (Harvard University)

Chair: Kadriye Ercikan (ETS Research Institute)

- **Is AI a Foundational Competency or Does AI Elevate Aspects of Competencies?**
Presenter(s): Michael Russell, Boston College
- **Don't Throw the BAIby Out with the Bathwater: Connecting FCEM to AI**
Presenter(s): Magdalen Beiting-Parrish
- **AI Competencies in Education and Credentialing: Aligning Expectations with Evolving Practice**
Presenter(s): Susan Davis-Becker, ACS Ventures, LLC
- **Technical AI Competencies as Measurement Competencies**
Presenter(s): Matthew Johnson, ETS Research Institute

Adaptive Testing, Assembly & Exposure Individual Paper Session

1:45 PM – 3:15 PM

Conference Center: King's Garden 4

- **Stochastic Constrained Test Assembly for AI-Enabled Assessment Systems**
Presenter(s): Alina von Davier, Duolingo
Modern assessment systems increasingly rely on dynamic item pools, continuous item generation, and on-demand test delivery, and challenge traditional approaches to test assembly. This paper introduces a stochastic constrained hybrid method for linear test assembly in AI-based testing environments that integrates utility-based stochastic sampling, reinforcement learning, and optimization-based constraint enforcement.
- **S2A3: Thompson Sampling and Stochastic Exposure Control for High-Stakes CATs**
Presenter(s): James Sharpnack, Lovelace
We introduce S2A3, a unified Bayesian framework for high-stakes computerized adaptive testing that eliminates separate item piloting. Thompson sampling routes uncertain items to informative test-takers while soft scoring attenuates their influence on ability estimates. Stochastic Sympon-Hetter exposure control ensures bank security. Validation on the Duolingo English Test confirms rapid calibration.

- **Assessment Use Guide Builder: Leveraging Use Cases to Generate Interpretive Assessment Materials**
Presenter(s): Yaning Cao; Nathan Dadey, The National Center for the Improvement of Educational Assessment
This study presents a Generative AI (GenAI) tool designed to create educator-facing interpretive materials, referred to as use guides, tailored to specific use cases. Through a four-step workflow, system prompt design, embedded guardrails, and three-round refinement, the tool produces context-specific, data-grounded interpretations and actionable recommendations.
- **Automated Pipeline for Scoring Numbers Reversed Items: Transcription-Based Reliability and Validity**
Presenter(s): Evelyn Johnson, Riverside Insights; Hsin-Ro Wei; Matthew Naveiras, Riverside Insights; Onur Demirkaya, Riverside Insights
Working memory assessment requires trained psychologists, creating barriers amid school-psychologist shortages. We developed an AWS SageMaker pipeline that transcribes and scores children's Numbers Reversed responses, achieving $\kappa = 0.77$ human-automated agreement. Transcription failure, not scoring logic, drove most errors. Full-scale results ($N \approx 13,000$) will be presented if accepted.
- **Fractional Information Gain: A Personalization Metric for Black-Box Predictive Models**
Presenter(s): S. Thomas Christie, Renaissance Philanthropy; Markus Hauru, Renaissance Philanthropy; Anna Rafferty, Carleton College
Personalization systems use student histories to select interventions. Predictive quality is quantified by ML metrics like AUC and accuracy, but we show these are misleading measures of personalization. To address this, we propose Fractional Information Gain, a metric analogous to psychometric reliability that quantifies the predictive capacity of individual student response histories.

Advancing Multimodal AI and Assessment System Innovation

Panel Session

1:45 PM – 3:15 PM

Conference Center: Commonwealth 2

Leading assessment solution providers discuss AI-powered technical product innovations that leverage extended principled assessment design to improve user-friendly, reliable, meaningful, valuable, and efficient assessment. Achievement Network, Carnegie Learning, Educational Collaboratory, Khan Academy, and New Meridian improved teacher and student perceptions, increased accessibility, and enhanced timely, instructionally relevant insights.

Moderator: Moderator: Eric Tucker (The Study Group)

- **Advancing Multimodal AI and Assessment System Innovation**
Presenter(s): Eric Tucker, The Study Group; Kristen Dicerbo, Khan Academy; Kristopher John; Christina Cipriano, Education Collaboratory, University of Massachusetts Amherst; Steven Ritter, Carnegie Learning; Lanette Trowery, Achievement Network

Clustering and Synthesis Techniques Using AI

Individual Paper Session

1:45 PM – 3:15 PM

Conference Center: Brigade

- **An Annotation Collection and Management System for AI-Based Scoring Systems**

Presenter(s): Janet Mee; David Canniere, National Board of Medical Examiners (NBME); Shirley Regan, NBME; Jianfang Du, NBME

The Annotation Collection Management System (ACMS) combines commercial and homegrown applications for collecting, validating, and managing annotation data for constructed-response scoring with AI. Attendees will learn how this human-in-the-loop workflow supports high-quality training data and model evaluation within an end-to-end ecosystem.

- **Rubricator: An AI Tool for Research-Aligned Rubric Design and Implementation**

Presenter(s): Sandra Liliانا Camargo Salamanca, University of Illinois Urbana-Champaign; Fabio Andres Parra-Martinez, Mississippi State University

Rubricator is an agentic AI tool that helps educators design and implement rubrics grounded in Rubric-Based Instruction (RBI) (Authors, 2026). The RBI framework is based on meta-analytic evidence of the effects of rubrics on student performance, self-efficacy, and self-regulation. Rubricator guides users through creation, alignment, and pedagogical implementation, using research-validated

- **AI-Assisted Thematic Analysis of Open-Ended Survey Responses**

Presenter(s): Kathryn Thompson, Pearson; Joshua Gilbert; Juste Mehrou, James Madison University; Jeffrey Steedle; Lissette Tolentino, University of Central Florida - College of Medicine; Doris Zahner, Council for Aid to Education

This demonstration showcases how generative AI can support thematic analysis of open-ended survey responses. Attendees will learn how to write effective prompts, evaluate AI-generated themes, and consider ways to integrate AI into qualitative analysis projects, using real feedback on the Standards for Educational and Psychological Testing as an applied example.

- **Beyond Latent Factors: Machine Learning Approaches to Response Clustering**

Presenter(s): Yi-Chen Wu, National Center on Educational Outcomes, University of Minnesota

This study explores the use of AI-based clustering methods to analyze survey responses and generate cluster-based feedback. Using UMAP for dimensionality reduction and K-means clustering, we identified moderate respondent groupings that were more represented than with traditional PCA. Findings suggest machine learning offers a promising alternative to factor-based survey analysis.

- **A Pipeline for Discovering Response-Pattern Constructs in Multiple-Choice Data for Feedback Generation**

Presenter(s): Joe Grochowalski, College Board; Claudia Ventura, College Board

Good feedback should target addressing misconceptions that lead to incorrect MCQ responses. Our pipeline detects these patterns automatically, then has a large language model (LLM) diagnose each as a misconception, test-taking habit, reading frame, or item flaw. Independent LLMs label each pattern, giving every diagnosis a reliability check.

Integrating Cognitive Models and AI to Enhance Assessment, Learning, and Instruction Coordinated Paper Session

1:45 PM – 3:15 PM

Conference Center: Smithfield/Birmingham

While significant research has explored cognitive models for learning, artificial intelligence presents the opportunity to reshape and expand their application and use. This coordinated session brings together five projects integrating AI with cognitive models to enhance assessment, educator practice, and student learning. The first paper examines emerging approaches for enabling AI systems to read and interpret learning maps—complex graphical models of knowledge and skill relationships—and explores how these representations can be leveraged for downstream learning and assessment applications. The second paper compares two modeling approaches for a learning progression-based assessment. The third paper investigates transformer based knowledge tracing models through an ablation study that clarifies how model components contribute to predicting student responses. The fourth paper introduces a machine learning tool that combines student data with underlying cognitive models to generate individualized recommendations for the next instructional targets students are most likely to master. The final paper describes a conversational agent designed to help educators accurately make sense of and use score reports. A discussant will synthesize themes across the five projects, highlighting implications for other AI-enabled assessments.

Discussant: Diego Zapata-Rivera (ETS)

Chair: Amy Clark (ATLAS: University of Kansas)

- **Developing and evaluating an AI agent to read and interpret learning maps**
Presenter(s): Dante Cisterna, ATLAS - University of Kansas; Jonathan Schuster, ATLAS - University of Kansas; Amber Samson, ATLAS - University of Kansas
- **Comparing Foundation and Supervised Vision Models for Learning Progression-Based Scientific Model Assessment**
Presenter(s): Leonora Kaldaras; Yasasvi Nalla, University of Houston
- **When Does Text Help Knowledge Tracing? Study of Language Model-Based Response Prediction**
Presenter(s): Jinnie Shin, University of Florida; Pauline Aguinalde, University of Florida
- **Next Skill Evaluation & Targeting via Map-Based Analytics**
Presenter(s): Amy Clark, ATLAS: University of Kansas; Jeffrey Hoover
- **Score Report Conversational Agent to Guide Educators in Interpreting Assessment Results**
Presenter(s): Brian Gane; Dante Cisterna, ATLAS - University of Kansas; Amy Clark, ATLAS: University of

Measurement Evidence for AI-Supported Assessment Design, Scoring, and Monitoring Coordinated Paper Session

1:45 PM – 3:15 PM

Conference Center: Grand Ballroom 1

This coordinated paper session examines how measurement methods can support the design, scoring, and monitoring of AI-supported assessments. The papers address the assessment lifecycle from generated content to operational score interpretation. One paper uses generalizability theory to evaluate LLM-generated multiple-choice science items across passages, random seeds, prompt templates, and item replicates. A second studies prompting and generation conditions that make synthetic student responses more realistic for automated-scoring data augmentation. A third uses multivariate generalizability theory to compare score dependability for AP Chinese speaking tasks completed with and without ChatGPT preparation support. A fourth develops confidence modeling for automated scoring, estimating when machine-assigned scores are more or less likely to agree with human scores and supporting more targeted monitoring of constructed-response scoring. A fifth introduces a regression-based approach to AES subgroup fairness, modeling score-distribution comparability and exact AES–human agreement across grades, traits, and demographic groups. Together, the papers argue that AI-supported assessment must be evaluated as a measurement system, not only as a model-performance problem. The session foregrounds variance decomposition, response realism, score dependability, confidence modeling, subgroup comparability, and validity evidence for defensible assessment use.

Discussant: Andrew Lan (University of Massachusetts, Amherst)

Chair: Edward Wolfe (University of Iowa)

- **Generalizability Theory for Evaluating Fine-Tuned LLMs in Automated Item Generation**
Presenter(s): Zhifei Li; Won-Chan Lee, University of Iowa
- **Prompting Strategies for Maximizing the Realism of Data-Augmented Responses**
Presenter(s): Bui Nhat Anh Vu, University of Iowa; Edward Wolfe, University of Iowa; Catherine Welch, University of Iowa
- **Evaluating Score Dependability in ChatGPT-Supported AP Chinese Speaking Tasks**
Presenter(s): Dan Song, The University of Iowa; Won-Chan Lee, University of Iowa
- **Confidence modeling using linguistic markers, score characteristics, and examinee characteristics**
Presenter(s): Edward Wolfe, University of Iowa; Justin Barber, Pearson
- **Joint Automated Scoring Fairness Modeling for Subgroups, Human Raters, and Automated Scoring**
Presenter(s): Yen Vo, The University of Iowa; Edward Wolfe, University of Iowa; Bui Nhat Anh Vu, University of Iowa

FULL SCHEDULE

WEDNESDAY, OCTOBER 7

PAD to PAAAD: Principled AI-Assisted Assessment Design

Coordinated Paper Session

1:45 PM – 3:15 PM

Conference Center: Heinz

AI can generate items from scant, or even naive prompting, so many assume we can discard the design knowledge expert item designers use in their craft and primarily rely on the LLM model to do the rest. We argue the opposite, that is, that the quality of items produced by generative AI will improve when we use the full power of our item design and development knowledge as inputs. Existing AI models cannot make weak assessment design strong nor help us improve the quality of items because the LLMs have been trained primarily on scores of items in the corpus, without critical design principles and context that expert item designers rely upon. We argue that PAD is essential when we are using AI to support test design and development. The session opens by making this case, and then we test our claim with three studies: the first examines which task model variables improve AI item generation; the second uses AI to study distractor plausibility as a predictor of mathematics item difficulty; the third predicts difficulty with both classical statistical models and modern machine learning. A discussant places the work in the broader context of AI and validity in educational measurement.

Discussant: Derek Briggs (University of Colorado - Boulder)

Chair: Kristen Huff (Curriculum Associates)

- **Why AI Demands More Principled Assessment Design, Not Less**
Presenter(s): Kristen Huff, Curriculum Associates
- **Which PAD Specifications Improve AI Item Generation?**
Presenter(s): Amanda Brice, Curriculum Associates; Catherine Feuerstein, Curriculum Associates
- **AI-Generated Distractor Plausibility Ratings as Predictors of Mathematics Item Difficulty**
Presenter(s): Jeffrey Steedle; Amanda Brice, Curriculum Associates; Catherine Feuerstein, Curriculum Associates
- **Predicting Item Difficulty: How PAD Features Shape Model Performance**
Presenter(s): Mihye Choi, Curriculum Associates

Responsible AI, Ethics & Transparency

Individual Paper Session

1:45 PM – 3:15 PM

Conference Center: King's Garden 5

- **How Residents Use Artificial Intelligence: Confidence, Concern, and the Case for Training**
Presenter(s): Linette Ross, National Board of Medical Examiners; Fen Fan, National Board of Medical Examiners
This study analyzes a national survey of medical sub-specialty residents (N = 3,662) on professional artificial-intelligence (AI) use. Most trainees use AI professionally yet receive little formal training. We describe use, verification, perceptions, and examine how confidence and concern co-occur with data-informed approaches to ethical AI use in clinical training.
- **The Analysis of Artificial Intelligence's Decade-Long Impact on Educational Measurement**
Presenter(s): Jiawei Xiong, Curriculum Associates
This study synthesizes international conference trends and peer-reviewed literature to chart AI's decade-long evolution in educational measurement. Quantitative method using systematic content analysis identify three developmental eras and four thematic domains, organized through an Efficiency–Enhancement–Transformation framework. Implications for responsible AI adoption and measurement-informed learning and instruction are discussed.

- Transformer-Aided Detection of Gaming in Constructed-Response English Language Assessments**
Presenter(s): Melanie Sharif, Pearson Education, Inc.; Scott Hellman, Macmillan Learning; Martha Bellows, Pearson; Susan Lottridge, Pearson
 This study compares classical machine learning and transformer-aided classification approaches for detecting gaming behavior in constructed-response items from an adult English language proficiency assessment. Using open-weight embedding models with dimensionality reduction and regularized logistic regression, preliminary results suggest transformer-derived features substantially improve detection accuracy beyond handcrafted feature baselines.
- Detecting Hallucinations in LLM-Generated Score Rationales**
Presenter(s): Josh Moskowitz, Acuity Insights; Cole Walsh, Acuity Insights
 Automated scoring systems which leverage large language models (LLMs) may be vulnerable to LLM-induced hallucinations. This paper aims to detect hallucinations in a LLM-as-judge system through a sampling based approach. This method successfully classified the majority of LLM outputs correctly. We discuss modifications to improve classification performance.
- Modeling Suspicious and Harmless Similarity in Multiple-Choice Test Security Investigations**
Presenter(s): Irina Grabovsky; Saed Rezayi, NBME
 Test-security analyses seek to distinguish suspicious shared information from ordinary agreement and harmless educational similarity. We present a work-in-progress generative latent-state framework for this task. Simulation results suggest that accurate modeling of expected agreement is more important than additional latent-state complexity and highlight directions for improving the current model.

Using AI for Support in the Writing Process

Individual Paper Session

1:45 PM – 3:15 PM

Conference Center: Duquense

- Generative Language Models for Argumentation Annotation**
Presenter(s): Kai North, Cambium Learning Group, Inc.
 This study integrates a generative language model (GLM) into an Automated Writing Evaluation system that identifies argumentative elements and language errors. Using parameter-efficient fine-tuning, the model accurately detects argumentative clauses, classifies relationships, and flags mistakes. Results show human-level performance while requiring substantially less computational power than larger models.
- Making Student Reasoning Visible: Comparing AI Approaches for Argument Structure Reconstruction**
Presenter(s): Man Jiang, University of Delaware
 This study introduces ArguGraph, a graph-based approach for reconstructing argument structures from student writing. Using expert-annotated essays, graph representations and AI approaches were compared in their ability to recover human-annotated argument structures. Preliminary findings suggest that preserving coordination structures improves reconstruction performance and supervised models outperform prompting-based large language models.
- From Revision to Transfer: Longitudinal Effects of LLM Writing Feedback**
Presenter(s): Yigal Attali, Duolingo; Bryan Smith, Arizona State University; Kai-Ling Lo, Duolingo; Andrew Runge, Duolingo; Jacqueline Church, Duolingo
 Prior automated writing feedback systems have shown limited evidence of transferable writing development. This pilot longitudinal study examined whether repeated interaction with process-oriented LLM feedback improved independent writing performance over time. Results suggest positive transfer effects across novel writing tasks, supporting the potential developmental role of LLM-mediated feedback.

- **Hierarchical Analytic Writing Feedback with Fine-Tuned Transformers**

Presenter(s): Michael Hemenway, Pearson; Justin Barber, Pearson; Martha Bellows, Pearson; Susan Lottridge, Pearson

We present a production system for hierarchical analytic writing feedback that uses frontier LLMs to generate training data for fine-tuned transformer models. Applied to Grade 5 opinion essays, the system scores 5 competency dimensions and 31 binary feedback codes, achieving agreement approaching inter-rater levels from training sets of 300–500 essays.

- **Beyond Prompting: Dialogic Act Modeling for AI Writing Tutors**

Presenter(s): Natasha Tetruyeva, eKidz; Yaroslav Nedashkovskyi, ekidz; Zelun Zhang, Technical University of Munich (TUM)

This demonstration showcases a conversational AI writing tutor guided by taxonomy-independent dialogic act modeling. Attendees will compare prompt-based and dialogic-act-guided tutoring interactions and learn how pedagogical dialogue strategies can be transferred across instructional settings without retraining, enabling more responsive and educationally effective AI tutors.

Using LLMS for Reading Passage Generation & Evaluation

Individual Paper Session

1:45 PM – 3:15 PM

Conference Center: Ft. Pitt

- **Automated Scoring of Oral Reading Fluency: An examination of validity evidence**

Presenter(s): Walter Leite, University of Florida; Krishna Sudeep Kumar, University of Florida; Jaiden Magnan, University of Florida; Stephanie Snidarich, University of Florida; Ahmad Rahimi, University of Florida; Zoey Liu, University of Florida

This study provides validity evidence for automated scoring of oral reading fluency (ORF) using off-the-shelf automatic speech recognition (ASR) models. A corpus of 320 audio recordings of 63 children collected with a digital literacy platform was used to estimate words read correct per minute (WCPM) with six model variants.

- **Beyond the Algorithm: Triangulating Evidence for ASR-Based Early Literacy Assessment**

Presenter(s): Aimee Boyd, Curriculum Associates; Mollie Headley, Curriculum Associates; Katelyn Louzeiro, Curriculum Associates; Yu Bai, Curriculum Associates; Karen Horowitz, Curriculum Associates

This pilot evaluates an automated speech recognition system for administering and scoring passage reading fluency in Grades 1–2 across authentic classrooms. Using multiple data sources including observations, surveys, interviews, and inter-rater analysis, we examine feasibility, student independence, and scoring accuracy. Preliminary results indicate strong AI–human agreement, informing continued development.

- **Generating Synthetic Student Reading Behavior for Passage Difficulty Estimation**

Presenter(s): Kuo Wang, Southern Methodist University; Akihito Kamata, Southern Methodist University

This study extends GEN-ORF by simulating student reading behavior, including reader ability, reading time, accuracy, prosodic pauses, and surprisal-informed timing variation. Preliminary results suggest that residualized silence and surprisal features jointly recover passage difficulty and timing parameters, supporting scalable passage difficulty estimation for additional passages before field testing.

- **Evaluation of AI-generated Passages for Reading Practice: An application of causal forests**

Presenter(s): Walter Leite, University of Florida; Qian Shen, University of Florida; Wonchae Lee; Sierra Evans; Sienna Perez, University of Florida; Wei Li, University of Florida; Jinnie Shin, University of Florida

This study presents a causal approach for evaluating automated generation of reading passages for early readers. It consists of estimating the average treatment effect of a candidate model, then conditional average treatment effects with causal forests. We apply it to evaluate a fine-tuned model for a digital literacy platform.

- **Design Choices in Encoder-Based IRT-3PL Parameter Prediction for Passage-Based Multiple-Choice Questions**

Presenter(s): Sungjin Nam, ACT Education Corp.

We study encoder-based prediction of IRT-3PL parameters for passage-based reading multiple-choice items. Experiments compare encoding strategies, cross-attention, joint versus per-parameter prediction, MSE versus CRPS losses, and passage-grouped batching. Separate-then-fuse encoding improves efficiency; joint CRPS training supports balanced recovery across discrimination, difficulty, and guessing in operational reading assessment item calibration workflows.

AI's Shifting Role in Measurement Science: Validation, Integration, and Prediction

Coordinated Paper Session

3:45 PM – 5:15 PM

Conference Center: King's Garden 4

As artificial intelligence reshapes measurement science in education, its role shifts across research contexts—from object of evaluation to measurement instrument, data source, and modeling engine. Ensuring psychometric rigor at each stage is the challenge this session addresses.

Treating AI as the object of measurement, Xiao and colleagues develop and validate an evaluation framework for LLM-based tutors, applying factor analysis to 192 candidate metrics and benchmarking LLM-as-judge scoring against human coders. Liu and colleagues deploy AI as a measurement instrument, harnessing automated speech recognition and NLP to derive student-level indicators of mathematical discourse quality from classroom audio, evaluating reliability and validity using IRT and generalizability theory. Shifting AI's role to data source, Starrett and colleagues integrate LLM-coded open-ended survey responses alongside closed-ended Likert items within SEM-based measurement models of teacher working conditions, comparing specifications for fit and discrimination of teacher leavers from stayers. Finally, Kelcey and colleagues embed AI into the modeling architecture itself, introducing targeted-after-generative (TAG) learning—combining variational autoencoders with targeted learning—for prediction and causal inference with AI-based assessments.

Together, the papers advance measurement science standards across the full arc—from evaluating AI itself, through using AI to derive and integrate indicators, to embedding AI in the modeling architecture.

Chair: Angela Starrett (University of South Carolina)

- **Developing an LLM Tutor Quality Evaluation Scale**

Presenter(s): Michael Xiao, University of Washington; Lief Esbenshade, University of Washington; Zewei Tian; Alex Liu, University of Washington; Min Sun, University of Washington

- **Using AI to Measure Student Mathematical Competencies from Classroom Discourse**

Presenter(s): Jing Liu, University of Maryland; Heather Hill, Harvard University; Michael Chrzan, Center for Educational Data Science and Innovation, University of Maryland

- **LLM-Coded Open-Ended Responses in Teacher Working Conditions Measurement and Retention Prediction**

Presenter(s): Angela Starrett, University of South Carolina; Katherine Masyn, University of South Carolina; Donna Coffman, University of South Carolina; Brian Cartiff, University of South Carolina; Beatrice Quiroz, University of South Carolina; Svetlana Dmitrieva, University of South Carolina

- **AI-Based Prediction and Causal Inference with AI-Based Assessments**

Presenter(s): Ben Kelcey, University of Cincinnati; Amota Ataneka, University of Cincinnati; Jean-Baptiste Habarurema, University of Cincinnati

Automated Item Generation & Item Writing

Individual Paper Session

3:45 PM – 5:15 PM

Conference Center: Ft. Pitt

- **Automated Item Evaluation: Predicting Item Acceptance and Rejection using LLM-Generated Critiques**

Presenter(s): Hotaka Maeda, Smarter Balanced; Yikai Lu, University of Minnesota Twin Cities

We trained transformer-based language models to predict whether items are acceptable for operational use using historical acceptance and rejection data. The raw item text classifier and a second classifier on Qwen3-generated item critiques were combined to improve performance. Model item rejection probabilities are evaluated based on reasons for rejection.

- **Automatic Item Generation Using a Multi-Agent Large Language Model Framework**

Presenter(s): Henry Makinde

A multi-agent LLM framework generated multiple-choice cardiovascular items for PAEA End of Rotation™ examinations across five clinical specialties. Of 333 items, 76 advanced to pilot testing at \$0.16 per item. Field-tested items showed difficulty and discrimination comparable to traditionally developed items, demonstrating psychometric viability.

- **Human-in-the-Loop Gemini Item Generation for Large-Scale MCQ Banks**

Presenter(s): Melchor Sánchez-Mendiola, National Autonomous University of Mexico (UNAM); Milton García-Lima, National Autonomous University of Mexico (UNAM); Elibidú Ortega-Sánchez, National Autonomous University of Mexico (UNAM); Manuel García-Minjares, National Autonomous University of Mexico (UNAM); Enrique Buzo-Casanova, National Autonomous University of Mexico (UNAM)

Using a human-in-the-loop workflow, we compared Gemini-generated and faculty-authored multiple-choice items for a large-scale UNAM item bank. Expert-review comments across four knowledge areas showed lower validation friction for AI-generated items in three areas, supporting AI as structural generator and faculty as final validators.

Building AI Equity Tools with Measurement Rigor: Four Developmental Cases

Coordinated Paper Session

3:45 PM – 5:15 PM

Conference Center: Brigade

Artificial intelligence holds promise for advancing equity in education, yet the field lacks shared standards for how AI tools built in service of equity should be developed, validated, and evaluated. This coordinated paper session argues that responsible development of such tools requires the same evidentiary standards applied to any educational instrument, including attention to construct validity, scoring transparency, iterative testing, and fairness. Drawing on four tools developed within a shared equity framework, this session presents developmental cases spanning distinct levels of education, from curriculum materials and evidence translation to measurement instruments and policy discourse. My Curriculum Partner applies AI-assisted analysis to examine whether curriculum materials create or remove barriers for students with learning differences. Sarphie Sense is a constrained, pipeline-based chatbot that supports navigation of a living systematic review evidence base. EquiFrame uses generative AI to identify and score disability language framing within educational measurement instruments. Crawl applies AI scoring to political leaders' public statements about disability, building a population-level discourse dataset. Across these four cases, researchers will discuss shared design principles including rubric development, human-AI scoring alignment, scope control, and backend validation, and invite dialogue about what measurement science demands of AI tools built in service of educational equity.

Discussant: Eric Tucker (The Study Group)

Chair: Christina Cipriano (Education Collaboratory, University of Massachusetts Amherst)

- **EquiFrame: An AI-based Program that Assesses Disability Language in Educational Measurement Instruments**
Presenter(s): Michael McCarthy, Education Collaboratory at Yale University
- **SarphieSense: A Pipeline-Based AI Chatbot for Navigating Living Systematic Review Evidence**
Presenter(s): Sophie Barnes, Yale University
- **My Curriculum Partner: An AI-Assisted Curriculum Audit and Language Analysis Tool**
Presenter(s): Christina Cipriano, Education Collaboratory, University of Massachusetts Amherst
- **Crawl: Holding Political Speech About Disability to a Measurable Standard at Scale**
Presenter(s): Michael McCarthy, Education Collaboratory at Yale University

Building Resilience Infrastructure for Digital Learning and Assessment

Coordinated Paper Session

3:45 PM – 5:15 PM

Conference Center: Commonwealth 1

AI-driven digital learning and assessment systems are inherently dynamic. While responsible AI (RAI) discussions often frame compliance as a set of fixed principles, the dynamic nature of these systems points to the need for resilience infrastructure: capabilities built into these systems to anticipate and respond to disruption. Resilience infrastructure protects RAI principles such as validity, reliability, and fairness through continuous monitoring and rapid response to unexpected disruption. This coordinated session asks: How do we build resilience infrastructure to sustain RAI principles in deployed digital learning and assessment systems? Drawing on real-world use cases, presenters discuss instabilities such as semantic variability in LLM replies, irregular test-taker behavior, drift in automated item generation output, prompt and model-version drift, and persona drift in conversational AI systems. For each, presenters discuss the resilience measures that address them. Presenters represent testing organizations (Duolingo English Test, Educational Testing Service, National Board of Medical Examiners) and industry (Amazon Science), with Kristen DiCerbo, Khan Academy's Chief Learning Officer, serving as discussant. The session emphasizes resilience infrastructure as essential to RAI practice in dynamic learning and assessment systems.

Discussant: Kristen DiCerbo (Khan Academy)

Chair: Jill Burstein (Duolingo)

- **Semantic Variability of LLM-Generated Replies Across LLMs: Implications for Designing Conversation-Based Assessment**
Presenter(s): Jiangang Hao, ETS
- **Resilience Infrastructure for Conversational AI-Based Educational Applications**
Presenter(s): Andrew Emerson, National Board of Medical Examiners; Keelan Evanini, NBME; Kevin Frome, National Board of Medical Examiners; Le Ha, Ho Chi Minh City University of Foreign Languages and Information Technology; Peter Baldwin, National Board of Medical Examiners
- **From Validation to Resilience: AI-assisted Enemy Item Identification in Operation**
Presenter(s): Ye Ma, Amazon Web Services
- **Responsible AI in Digital Assessment: Case Studies with the Duolingo English Test**
Presenter(s): Siyuan Chen, Duolingo, Inc.; Xiaowan Zhang, Duolingo; Andrew Runge, Duolingo; Jacqueline Church, Duolingo

Fairness and Reliability Individual Paper Session

3:45 PM – 5:15 PM

Conference Center: Smithfield/Birmingham

- **Predicting Race and Ethnicity DIF Using Generative AI in Large-Scale Assessment**

Presenter(s): Weiwei Cui, College Board; Yuefeng tang, College Board; Oliver Zhang, College Board; Jianshen (Cassie) Chen, College Board; Yuan Ge, College Board; Min Wang, College Board; Claudia Ventura, College Board

This study extends an AI-assisted DIF screening framework to race and ethnicity DIF evaluation in large-scale assessment development. Custom GPT models are trained on item text with empirical DIF labels for Asian, Black, and Hispanic groups. The framework supports proactive DIF risk screening, targeted item review, and fairness-aware item design.

- **On Promoting Individual-Level Fairness in Automated Scoring**

Presenter(s): Michael Fauss, ETS; Matthew Johnson, ETS Research Institute; Ikkyu Choi, ETS

We investigate the problem of promoting individual fairness in automated scoring. Three models are trained and evaluated under different individual fairness constraints. The models show improved scoring consistency on the test set, but at the cost of slightly reduced accuracy and a tendency to regress scores towards the mean.

- **When AI Graders Are Challenged: Reliability Under Student Appeals**

Presenter(s): Weixuan Xiao, Teachers College, Columbia University

Student appeals expose a vulnerability in AI grading: models may overturn correct decisions when confronted with persuasive but unsupported challenges. We evaluate three model families, four prompting strategies, and challenge strengths. Stronger challenges can increase false concessions. Structured multi-agent review reduces false concessions, but often corrects fewer initial errors.

- **Adversarial Response Testing for AI Scoring Robustness**

Presenter(s): YoungKoung Kim, College Board; Mike Owellen, College Board; Iev Khov, College Board; Sowjanya Ganameneni, College Board; David Jasuan, College Board; Deepa Abhilash, College Board; Altaf Gohar, College Board

This demonstration introduces an adversarial testing app that generates realistic but intentionally flawed student responses, scores them against uploaded rubrics, and reports whether AI scoring detects the embedded misconception. Attendees will learn how adversarial response generation can reveal rubric gaps, fluency bias, and reasoning/evidence failures before operational use.

- **Fairness Sentinel: An AI Agent for Early DIF Risk Screening**

Presenter(s): Jianshen (Cassie) Chen, College Board; Ion Drembliuga, College Board; Hunter Miller, College Board; Kyle Goetschius, College Board; Steven Hobbs, College Board; Weiwei Cui, College Board

We developed Fairness Sentinel, an AI-assisted triage agent that flags potential subgroup bias in draft items before field testing. Using labeled DIF patterns, it returns item-level risk classifications, feature-based rationales, and construct-preserving revision suggestions, supporting early fairness review, human judgment, model refinement, and validation.

Formative Assessment & Feedback**Individual Paper Session****3:45 PM – 5:15 PM****Conference Center: Heinz**

Evaluating Multi-Phase and Granular Strategies for Evidence-Anchored Automated Scoring

Presenter(s): Alessia Marigo, Wisconsin University - Madison; Laura Wright, UW Madison; Linda Malkin, Wisconsin University - Madison; Xin Xie, University of Wisconsin-Madison

Abstract:

While valid automated scoring tools exist, they do not always provide evidence, and if they do, it is inaccurate. Yet, this evidence is key for delivering truly personalized feedback. We address this critical problem by evaluating multi-phase scoring, rubric redesign, and granular clause parsing architectures for improving evidence extraction accuracy.

- **AI supported formative assessment in secondary education: early evidence from a pilot**

Presenter(s): *Leo Wilkinson; Laura Steighner, Colleague*

This paper presents early evidence from a pilot of AI supported formative assessment in UK secondary education. Drawing on classroom deployment across multiple schools, it reports qualitative feedback on learner level reporting and instructional usefulness, and discusses implications for transparency, classroom integration, and measurement informed design while formal impact evaluation is ongoing.

- **Semantic Similarity is Not Enough for Comparing LLM and Human Rater Rationales**

Presenter(s): *Julie Jung, Harvard University; Max Lu, Harvard Graduate School of Education; Dogus Darici, University of Münster; Emilia Bruegge*

Comparing LLM-human rater rationales using semantic similarity risks conflating textual proximity with evaluative agreement. We test whether embedding-based similarity reflects qualitative coding distinctions across rationale pairs in medical education. Similarity declined as rationale length differences grew and was less effective at distinguishing whether LLMs preserved the human's central claim.

- **LLM-Driven Gamified Narrative Reconstruction for Language Formative Quizzes**

Presenter(s): *Daijin Yang, Northeastern University; Sirui Wu, University of British Columbia; Amery Wu, University of British Columbia*

Language tests often lack engagement, which may reduce persistence in low-stakes settings. We propose an LLM workflow that reconstructs established language-test items into a gamified narrative while preserving core measurement elements. We evaluate the workflow through ablation test, expert review, learner testing, surveys, and psychometric analyses.

- **Is This a Test? Learner Experience and Control of the Formative–Summative Shift**

Presenter(s): *Lesley Reilly, WGU Labs; Kjorte Harra, WGU Labs; Megan Imundo, WGU Labs; Cory Hammon, WGU Labs; Lauren Zito; Catrina Nieser, WGU Labs; Betheny Gross, WGU Labs*

We ask how learners experience a covert shift between a formative and summative assessment within an AI-powered skills validation system, and whether letting learners (versus the system) control the shift decreases the efficiency of learning measurement. Because self-evaluation can be imprecise, mistimed learner-controlled shifts may have experiential and measurement tradeoffs.

IRT, Calibration & Psychometric Modeling

Individual Paper Session

3:45 PM – 5:15 PM

Conference Center: King's Garden 5

- **Validating an LLM Reliance Types Scale: Structure, Invariance, and Rasch Evidence**
Presenter(s): Shahin Hossain; Tukhbita Nawmi, University at Buffalo; Priyanka Saha, Lesley University
The LLM-RTS offers a theory-grounded, invariance-tested measure for identifying reliance profiles and targeting AI-literacy support.
- **CDM Q-Matrix Discovery with LLMs: Fusing Domain Knowledge with Empirical Evidence**
Presenter(s): Susu Zhang, University of Illinois Urbana-Champaign; V. N. Vimal Rao, University of Illinois Urbana-Champaign
We present a method for diagnostic assessment Q-matrix specification combining LLM input with model-based empirical validation. An LLM generates an initial Q-matrix from item content, and response data guide subsequent Q refinement. The approach integrates substantive rationale with empirical evidence to support scalable and measurement theory-grounded diagnostic assessment.
- **A Preliminary Semantic Screen for IRT Local Dependence**
Presenter(s): Qian Shen, University of Florida
This work-in-progress study examines whether sentence embeddings can flag item pairs at risk for local dependence in multidimensional IRT. Using a 50-item personality inventory, preliminary analyzes link semantic similarity to residual item-pair dependence. Future work will test robustness across larger samples, additional datasets, and alternative diagnostics.
- **Offloading the White Elephant: Assembling with AI-estimated IRT Parameters Instead of Pretesting**
Presenter(s): Anthony Shiver, Law School Admission Council; Aolin Xie
We evaluate an AI pipeline utilizing ensembled multimodal models to estimate 3PL Item Response Theory (IRT) parameters. In an assembly study for a large-scale admissions test, our AI-estimated parameters achieved comparable form yield, test information, and reliability to traditional pretesting.
- **How Reproducible Are LLM-Generated Q-Matrices? Evidence from Downstream Classification**
Presenter(s): Minkwon Kim, Teachers College, Columbia University
Researchers increasingly use LLMs to construct Q-matrices but focus mostly on single runs. Regenerating Q-matrices for three open-weight models on TIMSS mathematics items and fitting DINA, we find two runs of one model reclassify 61–88 percent of students. The instability is a sampling effect; disagreement with experts persists under deterministic decoding.

LLM Tutors

Individual Paper Session

3:45 PM – 5:15 PM

Conference Center: Grand Ballroom 1

- **Can Small Language Models Teach Math Through Socratic Dialogue?**

Presenter(s): Dizhi Zhang, Mengchen Li, Zichen Wang, The University of Pennsylvania Graduate School of Education; Ziqiao Zeng; Shruti Mehta, University of Pennsylvania; Talita Souza, Federal Institute of São Paulo (IFSP), Brazil and Institute of Mathematics and Computer Sciences of University of São Paulo (USP), Brazil; Seiji Isotani

This study evaluates whether three locally deployed small language models (LLaMA-3.2-1B, Qwen-2.5-0.5B, and Gemma-3-1B) can implement Socratic tutoring in Grade 6–8 mathematics. Using a rubric-based protocol across 72 sessions, results show all models struggled to sustain guided inquiry, with Gemma performing best overall.

- **Assessing Small Language Models as Decimal-Arithmetic Tutors: A Measurement Framework**

Presenter(s): Mai Vuong, University of Pennsylvania; Shahana Ahmadli, University of Pennsylvania; Minseok Kim, University of Pennsylvania; Michelle Zhou, University of Pennsylvania; Shruti Mehta, University of Pennsylvania; Talita Souza, Federal Institute of São Paulo (IFSP), Brazil and Institute of Mathematics and Computer Sciences of University of São Paulo (USP), Brazil; Seiji Isotani

Small language models are increasingly proposed for low-cost, private classroom use, yet their tutoring quality is rarely measured. We evaluate three sub-2B models tutoring decimal arithmetic across multi-turn interactions, using a literature-grounded rubric that separates mathematical validity from pedagogical appropriateness, and find pervasive, confidently delivered failures.

- **Can LLMs Judge Pedagogy? Assessing Conversational AI STEM Tutoring with AI-as-a-Judge**

Presenter(s): Xintian Gao, University of Florida; Qian Shen, University of Florida; Xin Li, University of Florida

Large language models are increasingly used as automated judges in education, yet their ability to score pedagogical quality in AI tutor responses to K-12 STEM student inquiries remains underexplored. This study evaluates whether two LLM-based scorers, Nemotron-3-Super-120B-A12B and GPT-OSS-120B, can approximate human judgment of Socratic STEM tutoring responses.

- **Beyond Volume: How Cognitive Fingerprints Predict Residual Gain in AI Tutoring**

Presenter(s): Nina Deng, Kaplan North America; Aleena Raj, Kaplan North America

This study evaluates an AI-based tutoring tool in a pre-licensure exam preparation product. Using residual gain modeling, elastic-net feature selection, and Gaussian Mixture clustering on LLM-derived cognitive fingerprints, the study identifies distinct learner profiles and shows that the cognitive depth of student–AI interaction, not interaction volume, drives measurable learning gain.

- **Automated Daily Monitoring of AI Tutor Quality with LLM Judges**

Presenter(s): Kevin Womack, Khan Academy

Evaluating AI tutors at scale can pose significant challenges: the velocity/volume of user interactions, product changes, and model updates outpace expert annotation and render point-in-time studies stale. Through this demonstration, attendees will learn to design an LLM-judge pipeline for continuous measurement that keeps pace with a rapidly evolving product landscape.

Using LLMs for Abnormal Response Detection**Individual Paper Session****3:45 PM – 5:15 PM****Conference Center: Duquense**

- **Forensic Item Clones: Psychometric Interventions Against External AI Assistance**

Presenter(s): Jing Huang, Purdue University; Hua-Hua Chang, Purdue University

Examinees increasingly consult generative AI during digital tests, threatening score validity. We introduce forensic item clones—construct-equivalent items hiding machine-targeted instructions that steer a consulting AI toward a specific wrong answer, converting statistical suspicion into deterministic behavioral evidence. We report cross-model injection-efficacy results and a planned ability-recovery simulation.

- **Detecting LLM-Assisted Exam Responses via Response Similarity Mixture Modeling**

Presenter(s): Aaron Myers, American Board of Internal Medicine

LLM use on assessments undermines the validity of inferences about examinees. This study introduces a response-similarity mixture model that operationalizes item responses as a latent combination of human ability and LLM-assisted responding. Simulations and empirical evaluations show the model estimates LLM-assisted response rates and distinguishes LLM from non-LLM resource use.

- **Exploring Asymmetric Academic Integrity Signals in Search- and LLM-Assisted Learning**

Presenter(s): Rahul Divekar, Bentley University

In a within-subject study, students learned via search and LLMs, then wrote essays without them. Analyses show no detectable difference across stylometric and perplexity features; AI detection probability is tool-dependent; plagiarism overlap higher after search. Results suggest asymmetric detection signals across learning pathways, with implications for interpreting academic integrity tools.

- **Who Keeps Up with Whom? Assessing GenAI Text Detection Across Version Updates**

Presenter(s): Ahnaf Chowdhury Niloy, Georgia State University, Georgia, United States of America.; Golnoush Haddadian, Georgia State University, Georgia, United States of America.; Min Kyu Kim, Georgia State University, Georgia, United States of America.

Artificial Intelligence use raises academic integrity concerns, yet institutions rely on detection tools of questionable reliability. Rapid version updates to chatbots, detectors, and humanization tools remain unaccounted for in detection research. Across 960 detection scores from two waves, the findings reveal detection outcomes are fundamentally reshaped by these parallel updates.

- **PsyMAS: A Human-in-the-Loop Multi-Agent System for Auditable Test Security**

Presenter(s): Jujia Li, University of Alabama; Qipeng Chen, University of Alabama - Tuscaloosa; Kaiwen Man, University of Alabama

PsyMAS is a human-in-the-loop multi-agent AI-assisted psychometric test security review system. It integrates deterministic forensic indices, and evidence records with AI interpretation and analyst adjudication. We will evaluate detection coverage, traceability, auditability, and careful reporting on semi-simulated assessment data, while ensuring AI does not make misconduct decisions.