Beyond Bots: Promoting and Measuring with Facilitative AI

2-hour Training Session
Monday, October 5, 2026 from 9:00 to 11:00 am EST.

Simon Cullen (Ph.D., Princeton)
Michael Strambler (Ph.D., UC Berkeley)

Brief 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.

Extended 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, commonground, 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.

I*nstallation/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.

Building the LAIR: A Framework for AI Literacy, Application, Interpretation, and Responsibility

2-hour Training Session
Monday, October 5, 2026 from 9:00 to 11:00 am EST.

Dr. Matt Meador, Learning AI Institute

Brief Abstract

Artificial intelligence adoption is outpacing AI literacy. This interactive workshop introduces the LAIR Framework—Literacy, Application, Interpretation, and Responsibility, as a practical model for helping educators, institutions, and organizations develop responsible AI users. Participants will leave with implementation tools, governance strategies, and assessment approaches that support ethical and effective AI integration.

Extended Description

Learning Objectives

1. Understand the four components of the LAIR Framework.
2. Evaluate AI outputs using structured interpretation methods.
3. Apply governance and responsibility principles to educational AI use.
4. Design AI literacy experiences for students, faculty, and workforce learners.
5. Develop implementation plans aligned to institutional goals.

Session Relevance

The rapid adoption of generative AI has created a gap between access and responsible use. While institutions increasingly deploy AI-powered tools, many lack structured frameworks that guide learners in understanding, applying, interpreting, and responsibly governing AI outputs. The Learning AI Institute developed LAIR as a practical framework designed to bridge this gap. Unlike tool-focused training, LAIR emphasizes transferable competencies that remain relevant regardless of which AI platform is used. The framework has informed AI literacy programming, workforce development initiatives, AI governance discussions, educational technology implementation, and responsible AI adoption planning.

Software Requirements
N/A

Best Practices for Integrating AI into Educational Policy and Decision-Making

2-hour Training Session
Monday, October 5, 2026 from 9:00 to 11:00 am EST.

Lukman Raimi, Universiti Brunei Darussalam

Brief Abstract

Artificial intelligence is transforming educational policy, assessment, and institutional decision-making. This interactive workshop equips educators, researchers, and policymakers with practical frameworks for responsible AI integration, focusing on governance, institutional readiness, ethical considerations, evidence-based decision-making, and educational effectiveness in AI-enabled learning environments.

Extended Description

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

1. Explain the opportunities and challenges of integrating AI into educational policy and decision-making.
2. Assess institutional readiness for AI adoption using practical diagnostic frameworks.
3. Apply principles of CREP as a template for responsible and ethical AI governance in educational settings.
4. Develop strategies to leverage AI to improve assessment, teaching, learning, and institutional effectiveness.
5. Design an institutional roadmap for AI integration that balances innovation, accountability, and educational quality.

Content and Relevance

Many institutions adopt AI without clear policies, governance, or guidelines, making it essential for educational leaders to develop practical strategies for responsible AI integration that maintain quality and stakeholder trust. This training offers evidence-based best practices for AI implementation in education. Based on research in higher education AI deployment, teaching, leadership, institutional readiness, and responsible governance, it provides practical advice for decision-makers aiming to incorporate AI effectively.

Participants will explore:

- Institutional readiness for AI adoption.
- Strategic leadership and change management.
- Ethical AI governance and risk management.
- AI applications in assessment, teaching, and learning.
- Data privacy, transparency, and accountability.
- Policy frameworks for sustainable AI implementation.

Special focus will be given to balancing innovation with measurement quality, educational impact, and ethical governance. Participants will review real-world cases of both successful and failed AI projects and extract practical lessons for their own institutions. This session directly supports the conference theme by showing how educational institutions can adopt AI while upholding high standards of validity, transparency, fairness, accountability, and teaching effectiveness.

Software and Hardware Requirements

- ChatGPT (Free or Paid)
- Google Gemini (Free)
- Microsoft Copilot (Optional)

Installation Requirements:

No installation is required.
Browser-based access only.

Cost:

- Free versions are sufficient.

Hardware:

- Laptop, tablet, or smartphone with internet connectivity.

How to generate scenario-based learning and assessment for post-secondary courses with GenAI

2-hour Training Session
Monday, October 5, 2026 from 9:00 to 11:00 am EST.

Zuowei Wang, ETS
Benjamin Pierce, Institute for Intelligent Systems

Brief 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.

Extended 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:

a) a shareable library of exemplary SBLA prototypes across diverse disciplines (e.g., psychology, interdisciplinary studies, health science, writing);
b) an Open-source, GenAI-driven Authoring System for use by instructors;
c) a Professional Support System such as this training;
d) a Multimodal Media Literacy Framework and associated Wiki; and
e) 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

Integrating Generative AI into R Workflows: From APIs to Shiny Apps

4-hour Training Session
Monday, October 5, 2026 from 1:00 to 5:00 pm EST.

Christopher Runyon, National Board of Medical Examiners

Brief 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.

Extended 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
    - 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.

Generative AI for Item Development: A Hands-On Workshop

4-hour Training Session
Monday, October 5, 2026 from 1:00 to 5:00 pm EST.

Henry Sanmi Makinde, University of North Carolina, Greensboro
Hope Oluwaseun Adegoke, University of North Carolina, Greensboro
Mubarak Mojoyinola, University of Iowa

Brief 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.

Extended 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.

Text Classification with Large Language Models: Pipelines, Fine-tuning, and Measurement Validity

4-hour Training Session
Monday, October 5, 2026 from 1:00 to 5:00 pm EST.

Brittney Hernandez, University of Connecticut
Kylie Anglin, University of Connecticut
Claudia Ventura, University of Connecticut

Brief 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.

Extended 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 labelled 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://bit.ly/aime-con or email brittney.hernandez@uconn.edu.

Writing an AI-Native Dissertation

4-hour Training Session
Monday, October 5, 2026 from 1:00 to 5:00 pm EST.

Damian Betebenner, Center for Assessment

Brief 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.

Extended 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.

Sponsored Training Session (*First 30 of registrants are free)*

How Should We Benchmark AI Tutors? Toward Valid and Impact-Oriented Evaluation

Sponsored 4-hour Training Session
Monday, October 5, 2026 from 1:00 to 5:00 pm EST.

Rene Kizilcec, Cornell – National Tutoring Observatory
Clayton Cohn, Cornell – National Tutoring Observatory
Josh Marland, Cornell – National Tutoring Observatory
Kirk Vanacore, Cornell – National Tutoring Observatory
Zhuqian Zhou, Cornell – National Tutoring Observatory

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, 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.

Schedule

Preferred format: Half-day (4-hour) workshop. The session will function as both a symposium on current AI tutor benchmarking work and a structured working session on future directions. It combines brief framing, invited benchmark-developer perspectives, hands-on critique and co-design, and facilitated discussion. Questions and discussion will be encouraged throughout. Participants will receive sample benchmark materials and access to a browser-based hosted environment; no installation or programming experience is required.

0:00–0:20 — Welcome, framing, and introductions
Organizers will introduce AI tutor benchmarking as a measurement problem: defining the construct of tutoring quality, identifying what a score is intended to support, and considering what evidence is needed to justify score interpretations. The organizers will briefly introduce the National Tutoring Observatory's work-in-progress impact-oriented benchmark and explain the workshop's collaborative purpose. Participants will share their interests and relevant work to identify common questions and seed future collaboration.

0:20–1:00 — Landscape of AI tutor benchmarking and measurement challenges
This session will provide an overview of current approaches, including benchmarks of pedagogical knowledge, response quality, student-error diagnosis, scaffolding, adaptive instructional decisions,

and multimodal tutoring. Presenters will highlight common design choices—tasks, reference responses, rubrics, expert ratings, student preferences, and LLM-as-a-judge systems—and the measurement questions they raise. The session will conclude with a short introduction to the critique rubric that participants will use later in the workshop.

1:00–1:45 — Invited perspectives: current benchmarks and future directions
Invited benchmark developers will present short cases describing their benchmark's target construct, task design, scoring approach, validation evidence, and limitations. Potential examples include TutorSim/TutorMoments, MathTutorBench, and the Pedagogy Benchmark. Discussion will focus on what each approach can validly support, what it leaves unmeasured, and priorities for future development.

1:45–2:00 — Break

2:00–3:00 — Small-group critique and co-design
Participants will work in facilitated groups using the critique rubric and sample materials from existing and work-in-progress benchmarks. Groups will select one design problem—for example, defining tutoring quality, representing student context, evaluating simulated interactions, designing a scoring rule, or linking behavioral indicators to learning outcomes. Each group will identify the proposed construct claim, potential validity threats, needed evidence, and a revised design or research plan.

3:00–3:35 — Group presentations and cross-group synthesis
Groups will briefly present their analysis and proposed revisions. Facilitators will synthesize recurring themes, disagreements, and unanswered questions to identify shared priorities for the field.

3:35–4:00 — Building the community of practice
Attendees will identify how the work relates to their own research or development efforts and define next steps for continued collaboration. Organizers will distribute the critique rubric, working benchmark materials, reading list, and repository links.

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

References

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 375–385). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445901

Lane, H. C., & Kageler, B. (2026). *CSTutorBench: Benchmarking small language models as tutors for block-based programming*. arXiv. https://doi.org/10.48550/arXiv.2607.05571

Lelièvre, M., Waldock, A., Liu, M., Valdés Aspillaga, N., Mackintosh, A., Ogando Portela, M. J., Lee, J., Atherton, P., Ince, R. A. A., & Garrod, O. G. B. (2025). *Benchmarking the pedagogical knowledge of large language models*. arXiv. https://doi.org/10.48550/arXiv.2506.18710

Liu, Y., Li, C., Zhang, T., Wang, M., Zhu, Q., Li, J., & Huang, H. (2025). *Discerning minds or generic tutors? Evaluating instructional guidance capabilities in Socratic LLMs*. arXiv. https://doi.org/10.48550/arXiv.2508.06583

Macina, J., Daheim, N., Hakimi, I., Kapur, M., Gurevych, I., & Sachan, M. (2025). MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 204–221). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.11

Maurya, K. K., Srivatsa, K. A., Petukhova, K., & Kochmar, E. (2025). Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 1234–1251). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.57

Otero, N., Druga, S., & Lan, A. (2025). A benchmark for math misconceptions: Bridging gaps in middle school algebra with AI-supported instruction. *Discover Education, 4*, Article 277. https://doi.org/10.1007/s44217-025-00742-w

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*. https://arxiv.org/abs/2111.15366

Srinivasa, R. S., Che, Z., Zhang, C. B. C., Mares, D., Hernandez, E., Park, J., Lee, D., Mangialardi, G., Ng, C., Hernandez Cardona, E.-Y., Gunjal, A., He, Y., Liu, B., & Xing, C. (2025). *TutorBench: A benchmark to assess tutoring capabilities of large language models*. arXiv. https://doi.org/10.48550/arXiv.2510.02663

Weitekamp, D., Siddiqui, M. N., & MacLellan, C. J. (2025). TutorGym: A testbed for evaluating AI agents as tutors and students. In *Artificial intelligence in education* (pp. 361–376). Springer. https://doi.org/10.1007/978-3-031-98420-4_26

Yang, T., Guo, S., Jia, M., Su, J., Liu, Y., Zhang, Z., & Jiang, M. (2026). MMTutorBench: The first multimodal benchmark for AI math tutoring. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 23310–23332). Association for Computational Linguistics. https://aclanthology.org/2026.acl-long.1068/