

2026 NCME AIME-Con Training Sessions

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

2026 NCME AIME-Con Training Sessions

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.

Measuring AI Adoption Readiness: A Neuroscience-Informed Diagnostic Framework

2-hour Training Session

Monday, October 5, 2026 from 9:00 to 11:00 am EST.

Micah Margolis, neurocollective

Brief Abstract

Most organizations track AI adoption through activity metrics: logins and tool usage. This session introduces a diagnostic framework grounded in neuroscience and organizational psychology for measuring adoption readiness and outcomes, closing the gap between AI deployment and validated behavioral change science.

Extended Description

This session introduces the neurocollective AI adoption measurement framework, which approaches the problem through three interlocking layers:

Diagnostic Assessment (the n.print instrument). The n.print is a behaviorally-anchored diagnostic that identifies individual AI adoption readiness across four profiles derived from construct families in organizational psychology, cognitive load theory, and behavioral economics. The instrument is designed to surface adoption barriers invisible to activity-based metrics: identity threat, workflow misalignment, change readiness, and cognitive style. Participants complete a sample diagnostic during the session and examine its construct design, scoring logic, and validity argument.

Outcome Measurement (n.score). The n.score framework operationalizes AI adoption outcomes at the workflow level, separating productivity effects from capability development, and individual adoption from organizational readiness. The framework draws explicitly on the logic of formative and summative assessment design, offering direct value to participants working on learning analytics and adaptive systems.

Certification Architecture (AIE L1-L4). The AIE certification framework levels AI adoption competency across four tiers modeled structurally after Six Sigma. This architecture raises questions directly relevant to measurement science: how do you define proficiency at the construct level, what constitutes evidence of advancement, and how do you ensure fairness across role, industry, and prior exposure?

Software Requirements

Software Needed: A web browser connected to the internet. The n.print diagnostic and the measurement design workshop activity will be delivered through a live, browser-based experience. No paper handouts will be used.

Installation/Payment Requirements: No installation required. No downloads. No cost to participants. Attendees need only a device with a modern web browser and an active internet connection. Participants may complete the diagnostic ahead of time if the conference organizers choose to distribute a link in advance, or it will be administered live during the session.

2026 NCME AIME-Con Training Sessions

Hardware Requirements: Any laptop, tablet, or mobile device with a modern web browser. No special hardware required.

2026 NCME AIME-Con Training Sessions

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

2026 NCME AIME-Con Training Sessions

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.

2026 NCME AIME-Con Training Sessions

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.

2026 NCME AIME-Con Training Sessions

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

2026 NCME AIME-Con Training Sessions

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:

2026 NCME AIME-Con Training Sessions

- 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.

2026 NCME AIME-Con Training Sessions

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 Mojinyinola, 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.

2026 NCME AIME-Con Training Sessions

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,

2026 NCME AIME-Con Training Sessions

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.

Software Needed

To be determined.

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

2026 NCME AIME-Con Training Sessions

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

2026 NCME AIME-Con Training Sessions

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.

2026 NCME AIME-Con Training Sessions

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

Open-Source Tools for De-Identifying, Transcribing, and AI-Assisted Coding of Educational Data

Sponsored 4-hour Training Session

Monday, October 5, 2026 from 1:00 to 5:00 pm EST.

Kirk Vanacore, Cornell, National Tutoring Observatory
Zhuqian Zhou, Cornell, National Tutoring Observatory
Rene Kizilcec, Cornell, National Tutoring Observatory

Brief Abstract

Measurement researchers increasingly work with large collections of discourse and multimodal data—classroom recordings, tutoring sessions, interviews, and constructed responses—that require human-quality coding at scale. This half-day workshop introduces a suite of open-source tools developed by the National Tutoring Observatory for responsibly processing and annotating such data. Participants will learn to (1) apply privacy-preserving, locally executable de-identification pipelines that remove personally identifiable information while retaining analytically meaningful content; (2) transform audio, video, and screen-capture data into structured, analyzable transcripts; and (3) configure and validate AI-assisted annotation using Sandpiper, a system that orchestrates multiple AI models to support scalable, auditable, human-in-the-loop coding. Measurement concerns are emphasized throughout: construct validity of coding schemes, human–AI agreement, reliability estimation, and disagreement-centered review. The intended audience includes researchers, graduate students, psychometricians, and practitioners who work with qualitative, discourse, or multimodal educational data. No programming experience is required; familiarity with qualitative coding or rubric-based scoring is helpful but not assumed. Attendees should bring their own laptops. No software installation is required in advance—tools will run in the browser or in provided environments—though optional setup instructions will be distributed before the session.

Extended Description

Relevance to the measurement field. The expansion of AI in educational measurement has made it feasible to score constructed responses, code classroom and tutoring discourse, and analyze multimodal interaction data at unprecedented scale. Yet the value of these methods depends on issues central to measurement: whether automated codes reflect the intended constructs, whether they agree with trained human raters, whether reliability and error can be characterized, and whether sensitive student data are protected throughout the pipeline. Human coding remains the gold standard but is a costly bottleneck; black-box AI coding risks invalid or unauditable inferences. This workshop addresses that gap by presenting a complete, open-source, human-in-the-loop workflow—from raw multimodal recordings to validated, analysis-ready coded datasets—that treats AI-assisted annotation as a measurement process rather than a shortcut. Codebooks are handled like scoring rubrics; model outputs are evaluated against human judgment; disagreement and uncertainty are surfaced for review rather than hidden.

What attendees will accomplish. After the session, attendees will be able to: de-identify educational discourse data using a high-precision open-source pipeline while auditing for over-

2026 NCME AIME-Con Training Sessions

redaction of analytically important content (e.g., mathematical expressions mistaken for PII); convert audio, video, screen, and log data into structured transcripts that preserve timing, modality, and links to source media; configure an annotation task in Sandpiper, including codebook representation, prompts, and orchestrated model roles for initial labeling, verification, and uncertainty flagging; evaluate AI-assisted annotations against human codes using agreement and reliability evidence; and export annotation artifacts suitable for downstream modeling. Attendees leave with reusable materials: a de-identification checklist, a workflow diagram, a template codebook, a sample Sandpiper configuration, and a validation checklist.

Presenter expertise. The organizing team builds the open-source infrastructure of the National Tutoring Observatory, a large-scale AI data-infrastructure project that increases researcher access to tutoring data. Team members developed the de-identification pipeline, the multimodal transcription workflow, and Sandpiper, and have published peer-reviewed and preprint work on orchestrated multi-model annotation of educational discourse and on optimizing LLM annotation through multi-agent orchestration.

Software Requirements

The de-identification and transcription tools are already released as open-source repositories. Sandpiper is in active development with a public open-source release scheduled well before the conference; a hosted workshop environment and pre-generated annotation outputs will be available as a fallback so all activities can be completed regardless. The session emphasizes practical application—configuring tasks, auditing outputs, and validating results—rather than syntax or mathematical formalism.