AIME Conference

Conference Program

ncme
NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

Thank you to our 2025 AIME-Con Sponsors:

PLATINUM SPONSORS

GOLD SPONSORS

SILVER SPONSORS

Cambium Assessment

CollegeBoard

edCount LLC
because all students count

Gates Foundation

NBME

SUPPORTERS

FOCALPOINT
Teaching & Learning. Measured. Managed

MI
MEASUREMENT
INCORPORATED

University of Pittsburgh LRDC

WestEd
WestEd.org

Schedule at a Glance

MONDAY, OCTOBER 27

10 a.m.–6 p.m.	Conference Check-In Open
Noon–2 p.m.	Two-Hour Trainings (Optional add-ons during registration)
2:30–6:30 p.m.	Four-Hour Trainings (Optional add-ons during registration)

TUESDAY, OCTOBER 28

7 a.m.–5 p.m.	Conference Check-In Open
8–9 a.m.	Keynote Session
9:15–10:45 a.m.	Session Block 1
11 a.m.–12:30 p.m.	Session Block 2
12:30–1:30 p.m.	Lunch (provided)
1:45–3:15 p.m.	Session Block 3
3:30–5 p.m.	Session Block 4
5–6:30 p.m.	Reception with Poster Sessions

WEDNESDAY, OCTOBER 29

7 a.m.–5 p.m.	Conference Check-In Open
8–9 a.m.	Keynote Session
9:15–10:45 a.m.	Session Block 5
11 a.m.–12:30 p.m.	Session Block 6
12:30–1:30 p.m.	Lunch (provided)
1:45–3:15 p.m.	Session Block 7
3:30–5 p.m.	Session Block 8

Monday

Two-Hour Trainings: 12:00–2:00pm

Benedum
Getting Started with LLM Evaluation: A Primer for Psychometricians

Birmingham
Creating Actionable Classroom Assessments: PLDs, Performance Tasks, and ChatGPT to the Rescue

Ft. Pitt
Designing for Variability: AI-Driven Innovation to Facilitate Formative Assessment and Personalize Learning

Heinz
Deep Learning Model for Unstructured Data in Educational Assessment

King's Garden 4-5
Designing and Evaluating Generative AI Simulations to Support Teacher Learning

Smithfield
Introduction to AI Scoring in Python

Break: 2:00–2:30pm

Four-Hour Trainings: 2:30–6:30pm

Benedum
Introduction to AI-based Automated Item Generation and Automated Scoring

Birmingham
Fundamentals of Generative AI for Item Development

Ft. Pitt
What you Need to Know to Unlock your Generative Potential

King's Garden 4-5
Using Generative AI For Item Construction: State-Of-The-Art and Practical Lessons

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

Tuesday

08:00 AM

King's Garden 4-5

Keynote - Liberty Munson (Microsoft)

09:00 AM - 09:15 AM Break

09:15 AM

Benedum

How to Improve LLM-Based Scoring Systems to Score More Diverse Student Work

- *Long context Automated Essay Scoring with Language Models* - Christopher Ormerod (Cambium Assessment)
- *Linguistic proficiency of humans and LLMs in Japanese: Effects of task demands and content* - Anastasia Smirnova (San Francisco State University)
- *Simulating Innovation: Using Large Language Models to Evaluate the Innovation Capacities Scale in Graduate Education* - Yun-Han Weng
- *Mathematical Computation and Reasoning Errors by Large Language Models* - Edith Graf (ETS)
- *Evaluating AI-Supported Cultural Narratives as a Multimodal Assessment for Language and Digital Literacy* - Elizabeth Falzone (Daemen University)

Birmingham
Exploring AI-Grading and Evaluation within the Testing Process

- *Automatic Grading of Student Work Using Simulated Rubric-Based Data and GenAI Models* - Yiyao Yang (Teachers College, Columbia University)
- *Evaluating Generative AI as Expert Assessors of Student Skills* - Joe Grochowalski (College Board)
- *genAI-Augmented Knowledge, Skills, and Ability (KSA) Profiles for Standard Setting* - Joe Grochowalski (College Board)
- *Enhancing Essay Scoring with GPT-2 Using Back Translation Techniques* - Aysegul Gunduz (University of Alberta)

Ft. Pitt
Leveraging LLMs for Novel Assessment Design & Delivery

- *Exploring AI-Enabled Test Practice, Affect, and Test Outcomes in Language Assessment* - Jill Burstein (Duolingo)
- *Head Orientation Reveals Differential Attention Patterns Between Questioning Types in Parent-Child Reading* - Boyuan Liu (Department of Educational Psychology, The Chinese University of Hong Kong)
- *Assessing AI skills: A washback point of view* - Meirav Attali (Fordham University)
- *High Leverage Opportunities to Transform Social and Emotional Competence Assessment with AI* - Christina Cipriano (Yale University)

Heinz
Rethinking Validity in AI-Based Automated Scoring: Interpretability, Faithfulness, and Measurement Principles in Educational Assessment

- *Evaluating the Consistency of Attribution Methods in Automated Short Answer Grading (ASAG) Systems* - Wallace Pinto (University of Florida)
- *Comparing Attention and Attribution Mechanisms in Transformer Models Through the Lens of Eye Gaze Data* - Lingchen Kong (University of Florida)
- *Development and Initial Validation of an Automated Stealth Scoring System of Oral Reading Fluency* - Walter Leite (University of Florida)
- *Reframing Validity Arguments for Automated Scoring Systems by Centering on True Scores* - Corinne Huggins-Manley

King's Garden 4-5
Improving Formative Assessment and Feedback using LLMs and Benchmark Datasets

- *Detecting Math Misconceptions: An AI Benchmark Dataset* - Bethany Rittle-Johnson (Vanderbilt University)
- *Automated Formative Assessment of Student-Drawn Science Models* - Mingyu Feng (WestEd)
- *Using Large Language Models to Refine the Q-Matrix for Cognitive Diagnosis* - Wenchao Ma (University of Minnesota)
- *AI-Generated Formative Practice and Feedback: Performance Benchmarks and Applications in Higher Education* - Rachel van Campenhout (VitalSource)
- *Psychmet- Measurement Foundational Competencies ChatBot* - Henry Makinde

Smithfield
Validating AI in Educational Testing: Challenges and Future Directions

- *Validating the Scorpion AI Assistant* - Chris Foster (Caveon)
- *Can AI generated rationale provide evidence that AI scores are valid?* - Daniel McCaffrey (ETS)
- *Item Quality Consideration in Automatic Item Generation* - Alexander Hoffman (AleDev Research & Consulting)
- *Comparing Evaluation Methods for Output from LLM-Powered Educational Chatbots* - Magdalen Beiting-Parrish (Federation of American Scientists)
- *Function Ascription, Use-plans and How to Validate AI Tools Used in Educational Testing* - Sergio Araneda (Caveon)

10:45 AM - 11:00 AM Break

11:00 AM

Benedum
Advances in Scalable, Accurate, and Reliable Automated Scoring

- *Efficient AES: Dimensionality Compression and Distillation in Transformer-based Models* - Yi Gui (The University of Iowa)
- *Improving Automated Scoring Accuracy through Synthetic Response Generation and Validation* Corey Palermo (Measurement Incorporated)
- *Exploration of Generative Large Language Models for Automated Scoring of Long Essays* - Haowei Hua (Princeton University)
- *Operational Alignment of Confidence-Based Flagging Methods in Automated Scoring* - Corey Palermo (Measurement Incorporated)
- *Efficacy of ad-hoc confidence measures and conformal prediction intervals in hybrid scoring* - Mark Beck (Measurement Incorporated)

Birmingham
Reliability and Validity of AI-Generated Content

- *Optimizing Reliability Scoring for ILSAs* - Ummugul Bezirhan (Boston College)
- *Evaluating AI Methods for Coding Spatial Language from Children's Naturalistic Conversations* - Qingzhou Shi (Northwestern University)
- *Examining decoding items using engine transcriptions and scoring in early literacy assessment* Zachary Schultz (Cambium Learning Group, Inc.)
- *Beyond Agreement: Rethinking Ground Truth in Educational AI Annotation* - Danielle Thomas (Carnegie Mellon University)

Ft. Pitt
Using LLMs for Teacher Education and Formative Assessment

- *Using Large Language Models to Analyze Students' Collaborative Argumentation in Classroom Discussions* - Nhat Tran (University of Pittsburgh)
- *Investigating Fine-Tuned AI Avatars to Support Teacher Learning in Simulated Math Instruction* - Hallie Parten (University of Virginia)
- *LLM-Human Alignment in Evaluating Teacher Questioning Practices: Beyond Ratings to Explanation* - Ruikun Hou (Technical University of Munich)
- *Measuring Teaching with LLMs* - Michael Hardy (Stanford University)
- *Towards assessing persistence in reading in young learners using pedagogical agents* - Beata Beigman Kelbanov (ETS)

Heinz
Employing LLMs to Craft Non-Cognitive and Novel Assessment

- *Automated Item Neutralization for Non-Cognitive Scales: A Large Language Model Approach to Reducing Social-Desirability Bias* - Sirui Wu (University of British Columbia)
- *Using LLM for rating Social Desirability* - FELIPE Valentini (Graduate School of Psychology, Universidade São Francisco)
- *Leveraging Large Language Models for Q-Matrix Construction in Linear Logistic Test Model* - Mubarak Mojoyinola
- *AI Enhanced Postural Assessments for Educational Measurement* - Fariha Hayat Salman (American University in Dubai)

King's Garden 4-5
Using LLMs to Predict DIF, Complexity, and Difficulty

- *A Digital Intelligence Framework for Predicting and Mitigating DIF in Large-Scale Assessment* - Weiwei Cui (College Board)
- *Predicting Text Complexity Using Keystroke-Derived Features* - Gulsah Gurkan (Pearson)
- *Using Generative AI to Develop a Common Metric in Item Response Theory* - Peter Baldwin (National Board of Medical Examiners)
- *Predicting differential item functioning with images and text: Multimodal AI model* - Hotaka Maeda (Smarter Balanced)

Smithfield
Building a K12 Education AI Infrastructure: Datasets on Math Teaching & Learning Learning Practices

- *What Good Tutors Do: Towards Measuring Tutoring Moves at Scale with the National Tutoring Observatory* - Rene Kizilecec (Cornell University)
- *Annotating Language Data for AI Measure Development in Education* - Dora Demszky (Stanford University)
- *Creating A High-Quality Multimodal Dataset of U.S. Mathematics Classrooms* - Jing Liu (University of Maryland)
- *Building a Secure Cloud Platform for AI Analysis of MET Classroom Videos* - Sandra Tang (University of Michigan)

12:30 PM - 01:45 PM Lunch Provided by AIME-Con
Commonwealth

01:45 PM

Benedum
Building Validity Evidence through Integrated AI Feature-level and Rubric Analysis Approaches

- *Using AWE to Measure Writing Growth Among Middle School ELs and Non-Els* - Joshua Wilson (University of Delaware)
- *The role of explainability in validity: How can XAI contribute to validation practice in automated scoring?* - Sarah Hughes
- *Using AI-Detected Writing Elements to Support the Validation of a New Writing Rubric* - Amy Burkhardt (Cambium Assessment)
- *The Impact of an NLP-Based Writing Tool on Student Writing* - Karthik Sairam (Cambium Assessment)

Birmingham
Using LLMs to Support and Evaluate Students Throughout the Writing Process

- *Evaluating the Validity of AI-Generated Writing Feedback for Students with Learning Disabilities* - Samantha Goldman
- *Evaluating the Reliability of Human–AI Collaborative Scoring of Written Arguments* - Noriko Takahashi (M.S. in Computational Linguistics, Montclair State University)
- *Revising with Generative AI to Support Writing and Revising Instruction* - Andrew Potter
- *Evaluating Measurement Invariance in AI and Human Scoring of Narrative Writing* - Ernest Amoateng (Western Michigan University)

Ft. Pitt
LLMs for Feedback and Formative Assessment

- *Implementation Considerations for Automated AI Grading of Student Work* - Zewei Tian
- *Generative AI in the K–12 Formative Assessment Process: Enhancing Feedback in the Classroom* - Mike Maksimchuk (Kent Intermediate School District)
- *Generative AI Teaching Simulations as Formative Assessment Tools within Preservice Teacher Preparation* - Jamie Mikeska (ETS)
- *Evaluating GenAI Feedback in Classroom Assessment: A Synthesis of Reviews* - Elie ChingYen Yu

Heinz
Using LLMs to Classify Learning Behaviors & Performance

- *Cognitive Engagement in GenAI Tutor Conversations: At-scale Measurement and Impact on Learning* - Kodi Weatherholtz (Khan Academy)
- *Using Whisper Embeddings for Audio-Only Latent Token Classification of Classroom Management Practices* - Wesley Morris
- *Numeric Information in Elementary School Texts Generated by LLMs vs Human Experts* - Anastasia Smirnova (San Francisco State University)
- *Addressing Few-Shot LLM Classification Instability Through Explanation-Augmented Distillation* - William Muntean (National Council of State Boards of Nursing)

King's Garden 4-5
Incorporating AI into Educational Measurement: Faculty, Research, Curriculum, and Universities

- *Advancing AI in Measurement at the University of Massachusetts Amherst* - Stephen Sireci (University of Massachusetts Amherst)
- *Advancing AI in Measurement at James Madison University* - Brian Leventhal (James Madison University)
- *Advancing AI in Measurement at the University of Maryland* - Hong Jiao (University of Maryland)
- *Advancing AI at Washington State University* - Shenghai Dai (Washington State University)

Smithfield
Generating and Evaluating Complex Items Using LLMs

- *Towards Reliable Generation of Clinical Chart Items: A Counterfactual Reasoning Approach with Large Language Models* - Jiaxuan Li (University of California Irvine)
- *Toward Automated Evaluation of AI-Generated Item Drafts in Clinical Assessment* - Tazin Afrin (NBME)
- *From theory to practice, high-stakes item development and review with Generative AI* - Marcus Walker (National Commission on Certification of Physician Assistants)
- *Generating Cognitively Equivalent Multiple-Choice Items from Constructed-Response Tasks Using AI* - Hyemin Park (UC Berkeley)
- *Automatic Item Generation for BEST Plus Assessment Using Large Language Model Prompting* Yage Guo (Center for Applied Linguistics)

03:15 PM - 03:30 PM Break

03:30 PM

Benedum
Comparing Human and AI Evaluation and Generation within the Test Development Process

- *Automated Evaluation of Standardized Patients with LLMs* - Andrew Emerson (National Board of Medical Examiners)
- *How Model Size, Temperature, and Prompt Style Affect LLM-Human Assessment Score Alignment* - Max Lu (Harvard University)
- *From Human Judgment to AI: Coding Student Reasoning in Spatiotemporal Tasks* - Qingzhou Shi (Northwestern University)
- *Validating AI Scoring of Constructed Responses with Cognitive Diagnosis Assessment Framework* - Hyunjoo Kim (University of Illinois Urbana-Champaign)

Birmingham
Using LLMs to Score and Evaluate Students' Open Responses

- *Comparative Study of Double Scoring Design for Measuring Mathematical Quality of Instruction* Jonathan Foster (University at Albany)
- *Evaluating LLM-Based Automated Essay Scoring: Accuracy, Fairness, and Validity* - Yue Huang (Measurement Incorporated)
- *Towards evaluating teacher discourse without task-specific fine-tuning data* - Beata Beigman Klebanov
- *Reliability thresholds and validity evidence for automated discussion scores* - Benjamin Pierce (University of Pittsburgh)

Ft. Pitt
LLMs for Standard Setting & Test Blueprint Design

- *AI-Based Classification of TIMSS Items for Framework Alignment* - Ummugul Bezirhan (Boston College)
- *Pre-trained Transformer Models for Standard-to-Standard Alignment Study* - Hye-Jeong Choi (HumRRO)
- *Comparing Human and AI Standard Setting Results: Are Standard Setting Panels Obsolete?* - Stephen Sireci (University of Massachusetts Amherst)
- *Calibrating Test Items and Creating Score Scales Using Paired Comparisons and AI* - Ketan (University of Massachusetts, Amherst)
- *AI-Enhanced Standard Setting: Leveraging AI to Support Human Experts* - Ernest Amoateng (Western Michigan University)

Heinz
Ethics, Fairness, and Validity in LLM-Based Systems for Educational Measurement

- *Undergraduate Students' Appraisals and Rationales of AI Fairness in Higher Education* - Victoria Delaney (San Diego State University)
- *Toward Responsible Use of AI in Education Research and Development* - Rachel Garrett (American Institutes for Research)
- *Implicit Biases in Large Vision–Language Models in Classroom Contexts* - Peter Baldwin (National Board of Medical Examiners)
- *AI-Augmented Validation: Transparent and Human-Centered Innovations for Early Childhood Measurement* - Supraja Narayanaswamy (Acelero Inc.)
- *Bias and Reliability in AI Safety Assessment: Multi-Facet Rasch Analysis of Human Moderators* Chunling Niu (The University of the Incarnate Word)

King's Garden 4-5
Debating AI's Future in Assessment: What's the Rush? What's the Risk?

- *Moderator:* Sarah Quesen (WestEd)
- *Panelist 2:* Damian Betebenner (Center for Assessment)
- *Panelist 3:* Kristen Dicerbo (Khan Academy)
- *Panelist 4:* Keelan Evanini (NBME)
- *Panelist 5:* Susan Lottridge (Cambium Assessment)

Smithfield
Predicting & Validiting IRT Parameters Using LLMs

- *Evaluating Difficulty Alignment of AI Generated Reading Passages for NAEP Assessment* - Qiwei He (Georgetown University)
- *Validating AI-Generated Test Items: A Study of Developer Judgments and NLP Measures* - Meltem Yumsek Akbaba (Ministry of National Education, Turkey)
- *Comparing item parameters of expert vs. AI-generated reading comprehension questions* - Zuowei Wang (Educational Testing Service)
- *Predicting Item Difficulty Using the Comparative Judgments of Large Language Models* - Eren Asena (Classic Learning Initiatives, LLC)

05:00 PM

King's Garden 1-2-3

Poster Session and Reception

1. *The AI Study Buddy: ChatGPT's Potential for Student Self-Assessment in Argumentative Writing* Tram-Anh Tran Nguyen (University of Massachusetts, Amherst)
2. *Enhancing Item Difficulty Prediction in Large-scale Assessment: Using Large Language Models* Mubarak Mojoyinola
3. *Comparing AI tools and Human Raters in Predicting Reading Item Difficulty* - Hongli Li (Georgia State University)
4. *Leveraging LLMs for Cognitive Skill Mapping in TIMSS Mathematics Assessment* - Ruchi Sachdeva (Pearson)
5. *Evaluating Creativity with Multimodal Large Language Models: Comparing AI and Expert Ratings* Haeju Lee
6. *Prompt Engineering for LLM-Generated Likert-Scale Survey Responses* - Nicole Bonge (University of Arkansas)
7. *Identifying Biases in Large Language Model Assessment of Linguistically Diverse Texts* - Lionel Meng (University of Wisconsin - Madison)
8. *Scaling Cognitive Diagnostics Across STEM* - Jayson Nissen (Montana State University)
9. *Patterns of Inquiry, Scaffolding, and Interaction Profiles in Learner-AI Collaborative Math Problem-Solving* - Zilong Pan (Lehigh University)
10. *Enhancing Reading Comprehension Through AI-Powered Real-Time Feedback Systems in Digital Learning Environments* - Chen Liu (UC Merced)
11. *AI-Powered Coding of Elementary Students' Small-Group Discussions about Text* - Carla Firetto (Arizona State University)
12. *LLM Open-Ended Assessment Scores Ignore Explicit Gendered Name Cues* - Kayden Stockdale (Virginia Tech)
13. *Automatic Diagnosis of Students' Use of Number Lines to Solve Fraction Problems* - Dake Zhang (Rutgers University)

Wednesday

08:00 AM

King's Garden 4-5

Keynote - Kenneth Koedinger (Carnegie Melon)

09:00 AM - 09:15 AM Break

9:15 – 10:15AM Sponsor Innovation Showcase: Curriculum Associates

King's Garden 1–3

R&D Priorities for AI in Educational Measurement and Testing

- *Moderator:* Derek Briggs (University of Colorado - Boulder)
- *Panelist 1:* Kristen Huff (Curriculum Associates)
- *Panelist 2:* Susan Lottridge (Cambium Assessment)
- *Panelist 3:* Britte Cheng (Menlo Education Research)
- *Panelist 4:* Ikkyu Choi (ETS)

09:15 AM

Benedum

Techniques for Improving and Strengthening LLM-Based Automated Scoring of Student Writing

- *Input Optimization for Automated Scoring in Reading Assessment* - Ji Yoon Jung (Boston College)
- *Develop a Generic Essay Scorer for Practice Writing Tests of Statewide Assessments* - Yi Gui (The University of Iowa)
- *From Entropy to Generalizability: Strengthening Automated Essay Scoring Reliability and Sustainability* - Yi Gui (The University of Iowa)
- *Explainable Writing Scores via Fine-grained, LLM-Generated Features* - James Bruno (Pearson)

Birmingham

Using LLMs to Support State and Federal Policy Making

- *AI as a Mind Partner: Cognitive Impact in Pakistan's Educational LandscapeG* - Hammad Javaid
- *Optimizing Opportunity: An AI-Driven Approach to Redistricting for Fairer School Funding* - Jordan Abbott (New America, Education Funding Equity Initiative)
- *Leveraging LLMs on Teacher Voices to Drive Retention Policies* - Svetlana Dmitrieva (University of South Carolina)
- *Assessing Science Teachers' Generative AI Literacy: A Multi-Method Instrument Development and Validation* - Ruiping Huang (University of Illinois Chicago)

Ft. Pitt

LLMs for Supporting Human Judgment and Evaluation in Educational Applications

- *Comparing Education Chatbot Evaluations: LLM-as-a-Judge vs. Human Raters* - Ting Zhang (American Institutes for Research)
- *Developing Performance Level Descriptors Using Large Language Models* - Jinah Choi (Edmentum)
- *Leveraging multi-AI agents for a teacher co-design* - Hongwen Guo (ETS Research Institute)
- *Controllable Student Proficiency Simulation via Guided Decoding* - Dongliang Guo (University of Virginia)

Heinz

Using LLMs as Conversational Partners for Novel Formative Assessment

- *Using Large Language Models for Formative Assessment of Young Adolescents' Revision Quality* Tianwen Li (University of Pittsburgh)
- *Evaluating the Impact of LLM-guided Reflection on Learning Outcomes with Interactive AI-Generated Educational Podcasts* - Vishnu Menon (Drexel University)
- *Talking to Learn: A SoTL Study of Generative AI-Facilitated Feynman Reviews* - Maddie Mattox (University of Virginia)
- *Fairness in Formative AI: Cognitive Complexity in Chatbot Questions Across Research Topics* - Alexandra Colbert (College Board)

Smithfield

Using LLMs to Improve Test Development

- *Generative AI in academic Publishing: Comparative Analysis of Five Publishers' Policies Introduction* - Aakash Kumar (Texas A&M University)
- *When Machines Mislead: Human Review of Erroneous AI Cheating Signals* - William Belzak (Duolingo)
- *Leveraging Fine-tuned Large Language Models in Item Parameter Prediction* - Suhwa Han (Cambium Asessment)
- *Automated search algorithm for optimal generalized linear mixed models (GLMMs)* - Miryeong Koo (University of Illinois at Urbana-Champaign)

10:45 AM - 11:00 AM Break

11:00 – 12:00pm Sponsor Innovation Showcase: Pearson
King's Garden 1–3

11:00 AM
Benedum
Using LLMs for Automated Scoring of Novel Student Work Products

- *Correcting Score Distribution in Automated Scoring using Reinforcement Learning* - Kai North (Cambium Learning Group, Inc.)
- *Using LLMs to identify features of personal and professional skills in an open-response situational judgment test* - Cole Walsh (Acuity Insights)
- *Comparison of AI and Human Scoring on A Visual Arts Assessment* - Ning Jiang (Measurement Incorporated)
- *Scoring Creativity at Scale: AI Evaluation of Student-Generated Metaphors* - Ricardo Primi (Universidade São Francisco)

Birmingham
The Diagnostic Automated Inference Scoring Engine (DAIS-E): Features-Based AI Scoring Research

- *Validity Evidence Supporting the Diagnostic Automated Inference Scoring Engine (DAIS-E)* - William Skorupski (Data Recognition Corporation)
- *Psychometric Consistency of Item Parameters Estimated from Human and AI Scoring* - Joseph Fitzpatrick (DRC)
- *Assessing Differential Item Functioning in Automated Essay Scoring Using DAIS-E* - Cassondra Griger (Data Recognition Corporation (DRC))
- *Latent Dimensions in DAIS-E Essay Features: Factor Analytic Evidence for Diagnostic Feedback* Karl Konz (Data Recognition Corp)

Ft. Pitt
Rating and Predicting Item Difficulty with LLMs

- *Predicting Item Difficulty for Pretest Items in Large-Scale Assessments* - YoungKoung Kim (The College Board)
- *Medical Item Difficulty Prediction Using Machine Learning* - Hope Adegoke (University of North Carolina)
- *Investigation into how a language model automatically rates items* - Jae Jun Jong
- *Simulating Rating Scale Responses with LLMs for Early-Stage Item Evaluation* - Onur Demirkaya (Riverside Insights)

King's Garden 4-5
Invited Session - Can AI Measure Up to Our Standards?

- *Moderator:* Kristen Huff (Curriculum Associates)
- *Panelist 1:* Qiwei He (Georgetown University)
- *Panelist 2:* Jill Burstein (Duolingo)
- *Panelist 3:* Kristen Dicerbo (Khan Academy)
- *Panelist 4:* Julia Rafal-Baer (ILO Group)

Smithfield
Automated Scoring Engine Training Innovations

- *When Does Active Learning Actually Help? Empirical Insights with Transformer-based Automated Scoring* - Justin Barber (Pearson)
- *Augmenting Training Samples with Responses Generated from an LLM in a Small-Sample Setting* - Mo Zhang (Educational Testing Service)
- *Calibrating Generative AI to Produce Realistic Essays for Data Augmentation* - Edward Wolfe (Pearson)
- *Exploring the Utilities of the Rationales from LLMs for Automated Essay Scoring* - Hong Jiao (University of Maryland)
- *Automated Essay Scoring Incorporating Annotations from Automated Feedback Systems* - Christopher Ormerod (Cambium Assessment)

12:30 PM - 01:45 PM Lunch Provided by AIME-Con
Commonwealth

1:45 – 2:45pm Sponsor Innovation Showcase: Duolingo
King's Garden 1–3
Accelerating Item Calibration and Bank Refresh with NLP-Based Item Features

- - Manqian Liao (Duolingo)

01:45 PM

Benedum

Examining Foundation Model Inferences in Terms of Fairness, Interpretability, and Robustness

- *Investigating Adversarial Robustness in LLM-based AES* - Renjith Ravindran (ETS)
- *On the Representation of Racial and Ethnic Subgroups in AI-Generated Texts: A Case Study in Automated Essay Scoring* - Akshay Badola (ETS)
- *Effects of Generation Model on Detecting AI-generated Essays in a Writing Test* - Jiyun Zu (ETS)
- *Exploring the Interpretability of AI-Generated Response Detection with Probing* - Ikkyu Choi (ETS)
- *A Fairness-Promoting Detection Objective With Applications in AI-Assisted Test Security* - Michael Fauss (ETS)

Birmingham

AI-Enhanced Item Evaluation: Content Alignment and Item Difficulty Modeling

- *Automated Item Content Alignment in Educational Assessment: A Systematic Review* - Nan Zhang (University of Maryland)
- *Text-Based Approaches to Item Alignment to Content Standards in Large-Scale Reading & Writing Tests* - Yanbin Fu (University of Maryland, College Park)
- *Automated Alignment of Math Items to Content Standards for Large-Scale Tests* - Qingshu Xu (University of Maryland, College Park)
- *A Systematic Review of Text-Based Approaches to Item Difficulty Modeling* - Sydney Peters (University of Maryland, College Park)
- *Item Difficulty Modeling Using Fine-Tuned Small and Large Language Models* - Ming Li (University of Maryland)

Ft. Pitt

Employing LLMs to Enhance the Test Design Process

- *Integrating AI and Human Expertise for Competency Mapping in Emerging Tech Sectors: A Hybrid Methodology for Workforce-Aligned Educational Measurement* - Maria Oliveri (Purdue University)
- *Scalable and Explainable AI with SQL-Augmented Retrieval* - Xinhui Maggie Xiong (ExamRoom AI)
- *AI-Driven Performance Assessment Design for Educators* - Alan Koenig (UCLA CRESST)
- *Dynamic Bayesian Item Response Model with Decomposition (D-BIRD): Modeling Cohort and Individual Learning Over Time* - Hansol Lee (Stanford University)

Heinz

Content Generation, Improvement, and Evaluation with LLMs

- *Using Open Source Language Models for Automated Passage Generation and Evaluation* - Alexander Kwako (Cambium Assessment)
- *Augmented Measurement Framework for Dynamic Validity and Reciprocal Human-AI Collaboration in Assessment* - Daniel Oyeniran (The University of Alabama)
- *SME comparison of PubMed RAG vs non-RAG items and fact-checking* - Marcus Walker (National Commission on Certification of Physician Assistants)
- *Evaluating Deep Learning and Transformer Models on SME and GenAI Items* - Joe Betts (National Council of State Boards of Nursing)

King's Garden 4-5

Measurement-Centered Approaches to Evaluate GenAI Outputs for Education

- *A Responsible Ecosystem Model for AI-Enabled Assessments* - Jill Burstein (Duolingo)
- *Beyond the Hint: Using Self-Critique to Constrain LLM Feedback in Conversation-Based Assessment* - Tyler Burleigh (Khan Academy)
- *Measurement-Informed Approaches to Evaluate GenAI Outputs for Education* - Michelle Barrett
- *GenAI Evidence Hub for Assessment Research: Synthesizing the State of the Art Research for Validity, Reliability and Fairness* - John Whitmer (Learning Data Insights)
- *Tests as Socio-Technical Artifacts: First Thoughts on its Implications for Validity Theory* - Sergio Araneda (Caveon)

Smithfield

Hybrid Automated/Human Scoring in Large Scale Assessment: The Texas Model

- *Hybrid Scoring in Texas: The State Perspective* - Chris Rozunick (TEA)
- *Hybrid Scoring in Texas: Automated Scoring Perspective* - Susan Lottridge (Cambium Assessment)
- *Hybrid Scoring in Texas: Content and Hand-scoring Perspective* - David Sanderson (Pearson)
- *Hybrid Scoring in Texas: The Psychometric Perspective* - Elizabeth Ayers-Wright (Cambium Assessment)
- *Hybrid Scoring in Texas: The Technical Advisory Committee Perspective* - Andrew Ho (Harvard University)

03:15 PM - 03:30 PM Break

03:30 PM

Benedum

LLM-Driven Interactive Assessments: Generation, Adaptation, and Scoring at Scale

- *Developing Multi-Modal Language Tasks with Computational Psychometrics* - Steven Nydick (Duolingo)
- *A generative AI-powered Adaptive and Interactive large-scale Speaking Assessment* - Yigal Attali (Duolingo)
- *Pre-Pilot Optimization of Conversation-Based Assessment Items Using Synthetic Response Data* - Tyler Burleigh (Khan Academy)
- *Measuring Student Understanding via Multi-Turn AI Conversations* - Jing Chen (Khan Academy)

Birmingham

From Theory to Practice: Generative AI Applications in Complex Educational Assessment

- *A Framework for Live Scoring Constructed Response Items with Commercial LLMs* - Scott Frohn (Khan Academy)
- *Toward more principled approaches for AI measurement of complex skills* - Peter Foltz (University of Colorado - Boulder)
- *Using Large Language Models to Grade Story Retell Tasks: Automated Assessment and Uncertainty Detection* - Owen Henkle (Rising Academies)
- *Higher Interest and Lower Error Rates for Peer-Authored Math Problems* - Kole Norberg (Carnegie Learning)
- *Using Generative AI for Item Development and Item Difficulty Prediction* - Sonya Powers (Edmentum)

Ft. Pitt

Using AI-Extracted Data to Improve Test Development and Analysis

- *Compare Several Supervised Machine Learning Methods in Detecting Aberrant Response Pattern* - Yi Lu (Federation of State Boards of Physical Therapy)
- *Exploring Generative Artificial Intelligence for Data Extraction in Single Case Design Meta-analysis* Yaosheng Lou (University at Albany, SUNY)
- *AI-Augmented Pretesting: Supplementing Traditional Pretesting with Simulated Approaches* - Brad Bolender (Finetune by Prometric)
- *Using recursively generated synthetic responses to validate item-level variational encoder latent representation* - Michael Chajewski (Pearson)

Heinz

Predicting & Generating Student Responses Using LLMs

- *Predicting and Evaluating Item Responses Using Machine Learning, Text Embeddings, and LLMs* Evelyn Johnson (Riverside Insights)
- *Keystroke Analysis in Digital Test Security: AI Approaches for Copy-Typing Detection and Cheating Ring Identification* - Chenhao Niu (Duolingo, Inc.)
- *Equating Forms Using Synthetic Data via LLM Respondents and DCM Models* - Sergio Araneda (Caveon)
- *Exploring the Psychometric Validity of AI-Generated Student Responses: A Study on Virtual Personas' Learning Motivation* - Huanxiao Wang
- *Simulated Students Aligned with Item Response Theory for Question Difficulty Prediction* - Andrew Lan & Christopher Ormerod (Cambium Assessment)

King's Garden 4-5

Co-Creating What Counts: Leading Across AI and Measurement to Reimagine Educational Assessment

- *Moderator:* Jill Burnstein (Duolingo)
- *Panelist 1:* Kristen Dicerbo (Khan Academy)
- *Panelist 2:* Laura Hamilton (American Institutes for Research)
- *Panelist 3:* Lee Becker (Pearson)
- *Panelist 4:* Nitin Madnani (Duolingo)
- *Panelist 5:* Susan Lottridge (Cambium Assessment)
- *Panelist 6:* Victoria Yaneva (NBME)

Smithfield

Human-AI Interactions to Improve Formative Assessment

- *LLM-Based Approaches for Detecting Gaming the System in Self-Explanation* - Jiayi (Joyce) Zhang (University of Pennsylvania)
- *Evaluating Generative AI as a Mentor Resource: Bias and Implementation Challenges* - Jimin Lee (Clark University)
- *Personalized Feedback and Learning Assessment in Work-Integrated Learning: An Explainable Recommender System Approach* - Jinnie Shin (University of Florida)
- *Mapping Co-Creative Knowledge Construction in Human-Robot Interaction through the ICAP Framework* - Janet Shufor Bih Epse Fofang (University of Pittsburgh)