

AIME-Con 2026

**Artificial Intelligence in Measurement and Education
Conference (AIME-Con)**

Volume 3: Coordinated Session Papers

October 5-7, 2026

The AIME-Con organizers gratefully acknowledge the support from the following sponsors.

Platinum



Gold



*ets research institute



Silver



Supporter







duolingo english test

The future of language assessment is here

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



Built on the latest language assessment science

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



englishtest.duolingo.com



Measure what matters

One connected learning story

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

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

Visit PearsonAssessments.com/navvy to learn more

800-627-7271 | PearsonAssessments.com

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

*ets
research
institute

AI research that empowers every learner

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

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



Learn more about our research on AI for assessment capabilities



Discover our AI research expertise at NCME AIME 2026:

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

Connecting Data to Instruction



 i-Ready Inform™

LEARN MORE



ADVANCING ASSESSMENT, SUPPORTING OPTIMAL CARE

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

innovationsinassessment.org



The AAMC Strengthens the World's Most Advanced Medical Care



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

Learn more at aamc.org.

26-050 (01/26)

AAMC.ORG

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



**Advancing
the Standard**
View our
Research





College Board Is a Proud Sponsor of AIME-Con

Join our engaging sessions to learn how we're advancing innovative and responsible use of AI in educational measurement.



©2026 National Council on Measurement in Education (NCME)

Order copies of this and other National Council on Measurement in Education (NCME) proceedings from:

National Council on Measurement in Education (NCME)
520 S. Walnut St. Box 2388
Bloomington, IN 47402
USA
Tel: +1-812-245-8096
ncme@ncme.org

ISBN 979-8-9983004-2-4

Preface



Introduction

The second NCME-sponsored Artificial Intelligence in Measurement and Education Conference (AIME-Con) brought together an interdisciplinary community of experts working at the intersection of artificial intelligence (AI), educational measurement, assessment, natural language processing, learning analytics, and technological development. As AI continues to transform education and assessment practices, this conference provided a critical platform for fostering cross-disciplinary dialogue, sharing cutting-edge research, and exploring the technical, ethical, and practical implications of AI-driven innovations in measurement and education.

Conference Theme - Measurement Science in AI-Integrated Assessment and Pedagogy

The second NCME-sponsored Artificial Intelligence in Measurement and Education Conference (AIME-Con) focused on illustrating ways that artificial intelligence is reshaping assessment, feedback, and pedagogy, while underscoring that these innovations must be anchored in rigorous measurement science. As AI systems expand the possibilities for scoring, item generation, adaptive tasks, and personalized instructional support, they simultaneously create new demands for evidence about what these systems measure, how well they measure it, and how they affect learning. The conference theme “Measurement Science in AI-Integrated Assessment and Pedagogy” frames psychometric principles—validity, reliability, fairness, and instructional effectiveness—as the foundation for responsible AI design and evaluation. It highlights the necessity of integrating measurement theory, data science, and learning research to guide the development and use of AI tools. The theme signals a commitment to advancing not only the capabilities of AI but also the evidentiary standards required to ensure these technologies enhance education.

The National Council on Measurement in Education

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

NCME Special Interest Group on Artificial Intelligence in Measurement and Education

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

Proposal Requirements and Review Process for Coordinated Session Papers

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

- **Relevance and community impact:** pertinence to the conference theme and AI in measurement and education community broadly, and potential contribution to current discussions and challenges in the field.
- **Scholarly or Practical Significance:** scholarly merit or practical importance of the work, and potential impact on theory, practice, or policy.
- **Methodological rigor:** coherence and appropriateness of the proposed methods, techniques, and approaches; and soundness of the overall research design.
- **Strength and trustworthiness of the results and conclusions:** whether the proposed analysis and interpretation methods yielded trustworthy results and conclusions that contribute valuable knowledge to the field.

For the purposes of this conference, “AI” was defined broadly to include rule-based methods, machine learning, natural language processing, and generative AI/large language models. Reviewers provided constructive feedback and overall recommendations to ensure that accepted sessions reflected both scholarly merit and practical value to the AI in measurement and education community.

Organizing Committee

NCME Leadership

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

Conference Chairs

Joshua Wilson, Ph.D., University of Delaware
Christopher Ormerod, Ph.D., College Board
Magdalen Beiting-Parrish, Ph.D., EdAlfy, CUNY Graduate Center

Proceedings Chair

Nitin Madnani, Ph.D., Duolingo

Program Chairs

Joshua Wilson, Ph.D., University of Delaware
Christopher Ormerod, Ph.D., College Board
Magdalen Beiting-Parrish, Ph.D., EdAlfy, CUNY Graduate Center
Polina Harik, Ph.D., NBME

Program Committee

Program Chairs

Joshua Wilson, Ph.D., University of Delaware
Christopher Ormerod, Ph.D., College Board
Magdalen Beiting-Parrish, Ph.D., EdAlfy, CUNY Graduate Center
Polina Harik, Ph.D., NBME

Reviewers

Amy Adair, ETS
Beata Beigman Kelbanov, ETS
Magdalen Beiting-Parrish, EdAlfy, CUNY Graduate Center
Martha Bellows, Pearson
Aimee Boyd, Curriculum Associates
Christina Cipriano, Education Collaboratory, University of Massachusetts Amherst
Alex Codoreanu, Australian Council for Educational Research
Kristen Dicerbo, Khan Academy
Polina Harik, NBME
Qiwei He, Georgetown University
Matthew Johnson, ETS Research Institute
Hongli Li, Georgia State University
Xiaoran Li, Cognality Learning
Jing Liu, University of Maryland
Ruitao Liu, ACT
Ru Lu, Educational Testing Service
Mike Maksimchuk, Michigan Virtual
Joshua McGrane, University of Melbourne
Jamie Mikeska, ETS
Sebastian Nastuta, Pearson Education UK
Maria Oliveri, Purdue University
Christopher Ormerod, College Board
Rebecca Swinburne Romine, University of Kansas
Joshua Wilson, University of Delaware
Mo Zhang, Educational Testing Service

Table of Contents

<i>Multi-Agent LLM Annotation and Scoring for Training Fine-Grained K-12 Writing-Feedback Models</i> Justin O Barber, Michael P Hemenway, Martha Bellows and Susan Lottridge	1
<i>LLM-Based Rubric Refinement in Multi-Agent Automated Scoring</i> Alexander Kwako, Cristina Everett and Harry Wang	9
<i>Beyond Agreement: Operational Monitoring of K–12 Automated Scoring</i> Jing Ma, Edward W Wolfe and Anthony D Fina	23
<i>Detecting Invalid Responses in Automated Essay Scoring with Fine-Tuned Large Language Models</i> YoungKoung Kim and Christopher Ormerod	31
<i>Beyond Clean Agreement: A Stability-Bounded Evaluation of Compact AES Distillation</i> Yi Gui	39
<i>From Efficiency to Sufficiency: A Measurement-Informed Framework for Cost-Aware Automated Essay Scoring</i> Yi Gui	49
<i>Generalizability Theory for Evaluating Fine-Tuned LLMs in Automated Item Generation</i> Zhifei Li and Won-Chan Lee	59
<i>Evaluating Score Dependability in ChatGPT-Supported AP Chinese Speaking Tasks</i> Dan Song and Won-Chan Lee	70
<i>From Validation to Resilience: Sustaining Measurement Quality in AI-Assisted Enemy Item Identification</i> Ye Ma	76
<i>Semantic Variability of LLM-Generated Replies Across LLMs: Implications for Designing Conversation-Based Assessment</i> Jiangang Hao	82
<i>Resilience Infrastructure for Conversational AI-Based Educational Applications</i> Andrew Emerson, Keelan Evanini, Kevin Frome, Le An Ha and Peter Baldwin	89
<i>Responsible AI in the Duolingo English Test: Case Studies with Automatic Item Creation and Session-Level Quality Monitoring</i> Siyuan Marco Chen, Xiaowan Zhang, Andrew Runge and Jacqueline Church	101
<i>Developing and evaluating an AI agent to read and interpret learning maps</i> Dante Cisterna, Jonathan Schuster and Amber Samson	110
<i>Learning Progression-Guided Scientific Model Assessment Using Foundation and Supervised Vision Models</i> Leonora Kaldaras and Yasasvi Nalla	118
<i>Developing an LLM Tutor Quality Evaluation Scale</i> Michael Xiao, Zewei Tian, Alex Liu, Lief Esbenshade and Min Sun	124
<i>When Language Becomes a Shortcut in Automated Short-Answer Scoring</i> Xiaomeng Xiong, Corinne Huggins-Manley, Jinnie Shin and Christian Grant	146

<i>Disentangling Severity and Centrality: Rater Effects in LLM-Based Automated Short-Answer Scoring</i> Xiaomeng Xiong, Corinne Huggins-Manley and Jinnie Shin	164
<i>From Expert Approval to Validity Evidence: Evaluating AI-Generated Assessment Items</i> Chris Meador	180
<i>Evaluating an AI-Item Generation Tool for NAEP Reading</i> Cara Marie DiMeglio	189
<i>Evaluating an AI-Item Generation Tool for NAEP Mathematics</i> Cheryl Van Ness, Ilona Minchuk and Karen Parker	196
<i>Evaluating an AI-Item Generation Tool for NAEP Science</i> Carolina Safar, Mitchell Price and Jeffery Ackley	203
<i>Evidence-Centered Design for AI-Driven Automated Item Generation in Applied Mathematics</i> Edith Aurora Graf, Maria Elena Oliveri, Giulia Oliveri and Emily Proctor	210
<i>Human–AI Collaboration in Educational Measurement: Transforming Assessment Data into Educational Action</i> Laura C. Egan and Judy H. Tang	219
<i>Scaling Human-AI Collaboration: Translating Multi-Source Assessment Data into Formative Learner Profiles</i> Hongwen Guo, Matthew S Johnson, Luis Saldivia, Michelle Worthington, Jeremy Lee and Kadriye Ercikan	225
<i>Human–Machine Learning for Large-Scale Qualitative Coding in Educational Measurement</i> Judy H. Tang, Tom Krenzke, Jin Hui Xu and Karen Lo	236
<i>Rubric-Aligned Generative-AI Features as Supplementary Predictors in Feature-Based Automated Essay Scoring</i> Yue Huang, Duanli Yan and Corey Palermo	241
<i>Latent Factors of Discourse Type and Effectiveness in PERSUADE 2.0 for AES</i> Yi Gui	251
<i>Rethinking Validity in Educational Assessment when AI Co-Produces Performance</i> Xiaoran Li and Tanesia Beverly	260
<i>Programmatic Tiered Prompting for LLM Generation of ELA & Math Items</i> David Whitecomb, Marjorie Wine and Alexander Hoffman	267
<i>The Scoring Paradox: Multi-Agent Architectures for Unbiased Feedback Optimization</i> Okan Bulut, Bin Tan, Elisabetta Mazzullo and Cole Walsh	274
<i>Can LLMs detect item flaws in the questions they have generated?</i> Guher Gorgun	284
<i>Automated Evaluation of Mathematical Equivalence Between Personalized and Standard Word Problems</i> Burcu Arslan, Ikkyu Choi, Jesse R. Sparks, Reginald M. Gooch, Candace Walkington and Matthew L. Bernacki	291
<i>Next Skill Evaluation & Targeting via Map-Based Analytics</i> Jeff Hoover and Amy Clark	306

<i>Score Report Conversational Agent to Guide Educators in Interpreting Assessment Results</i>	
Brian D. Gane, Dante I. Cisterna and Amy K. Clark	312
<i>Automated Generation and Scoring of Maze Reading Comprehension Assessments</i>	
Hatice Kubra Karakis, Walter Leite, Logan Scott, Xinyi Tai and Akihito Kamata	320