

AIME-Con 2025

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

Volume 1: Full Papers

October 27-29, 2025

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

Platinum



Gold

Curriculum Associates



*ets research institute

Silver



Gates Foundation



Supporters

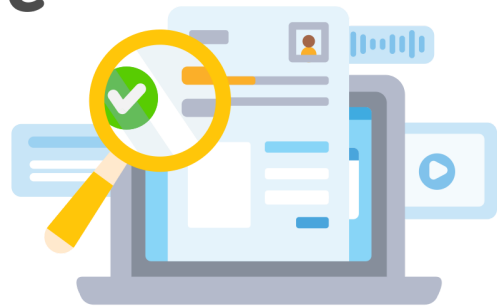




duolingo english test

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 5,800 programs worldwide



englishtest.duolingo.com



Evidence-based approach to AI in Measurement & Learning

At the intersection of artificial intelligence and educational measurement, Pearson stands as your trusted partner—delivering clarity, confidence, and innovation in every assessment moment.

Why Pearson?

- **AI-Enhanced Accuracy:** Using automated scoring and predictive analytics to provide insights that are accurate, fair, and timely.
- **Future-Ready Solutions:** Platforms that evolve with policy, pedagogy, and technology.
- **Personalized Learning Journeys:** Multi-lingual access and adaptive item generation to support each student's unique growth trajectory.
- **Ethical AI Practices:** Commitment to data security, transparency, explainability, and bias mitigation.
- **Collaborative Innovation:** Partnering with educators, researchers, and technologists to shape the future of assessment.

Human-Centric AI	Pearson believes AI's highest purpose is to elevate and empower human capabilities.
Assessment as a Learning Continuum	We reimagine assessments not as endpoints, but as integral parts of the learning journey.
AI as an Environment	Pearson is exploring how this shift impacts our approach to assessment—ensuring our tools are adaptive and future-ready.
Balancing Vision and Capabilities	We deliver reliable solutions today while building toward the future of AI in education.

The future of *i-Ready* Assessment is invisible.

- Voice technology is coming to *i-Ready* Literacy Tasks
- Built to hear students' voices of all accents and dialects
- Creating the best possible solution by collaboratively learning with teachers in the classroom



Learn more about our vision for the future



*ets research institute

Shaping the Future of AI in Assessment

ETS advances responsible AI research to promote fairness, trust, and innovation.

As AI transforms education, ETS brings decades of expertise to ensure that new solutions are not only powerful, but also valid, equitable, and transparent. Our work is driving the next-generation of measurement science, standing at the intersection of AI, learning, and assessment.

Highlights from ETS research at NCME AIME 2025:

- Investigating racial and ethnic subgroup representation in automated essay scoring
- Using generative AI teaching simulations to support teacher training
- Designing fairness-promoting, automated fraud detection systems
- Validating AI generated scoring rationales

REVIEW OUR
GUIDELINES FOR
RESPONSIBLE AI→



Advancing Assessment with AI

Grounded in science and responsible best practices, we use AI to enhance how we measure what students know and can do.

● **19 states**
we serve use hybrid scoring

24M essays & short answers
auto-scored by our AI engines

2M verbal responses
auto-scored by our AI engines

More AI-Powered Features - Coming Soon!

- WriteOn with Cambi
- Item Parameter Estimation
- Cheating Analysis
- Teacher Authoring with AI passage generation
- Hotline for student-at-risk work detection



Data reflects the 2024-2025 academic year



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.



edCount is pleased to sponsor 2025 NCME AIME-Con

*Over 20 years of service to
students and educators!*



Our Belief Statement

Every individual brings unique experiences, skill sets, and perspectives that work to advance our purpose: continuously improving the quality, fairness, and accessibility of education for all students.

Our Services

- Assessment Design, Development, and Evaluation
- Instructional Systems and Capacity Building
- Policy Analysis and Technical Assistance



www.edCount.com

(202) 895-1502 | info@edCount.com



www.NBME.org

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.



©2025 National Council on Measurement in Education (NCME)

Order copies of this and other 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 979-8-218-84228-4

Preface



Introduction

The inaugural 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. By bringing together experts from varied domains, the conference fostered a rich exchange of knowledge to enhance the collective understanding of AI's impact on educational measurement and evaluation.

Conference Theme - Innovation and Evidence: Shaping the Future of AI in Educational Measurement

The NCME-Sponsored AIME-Con focused on how rigorous measurement standards and innovative AI applications can work together to transform education. With sessions spanning summative large-scale assessment, formative classroom assessment, automated feedback, and informal learning tools, this conference fostered both the advancement and evaluation of AI technologies that are effective, reliable, and fair.

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 Full Papers

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

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

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

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

Amy Hendrickson, Ph.D. (President)
Rich Patz, Ph.D. (Executive Director)

Conference Chairs

Joshua Wilson, University of Delaware
Christopher Ormerod, Cambium Assessment
Magdalen Beiting Parrish, Federation of American Scientists

Proceedings Chair

Nitin Madnani, Duolingo

Proceedings Committee

Jill Burstein, Duolingo
Polina Harik, NBME

Program Committee

Conference Chairs

Joshua Wilson, University of Delaware
Christopher Ormerod, Cambium Assessment
Magdalen Beiting Parrish, Federation of American Scientists

Reviewers

Ketan , University of Massachusetts, Amherst
Hope Adegoke, University of North Carolina
Tazin Afrin, NBME
Ernest Amoateng, Western Michigan University
Kylie Anglin, University of Connecticut
Sergio Araneda, Caveon
Meirav Attali, Fordham University
Nurseit Baizhanov
Lee Becker, Pearson
Beata Beigman Klebanov, ETS
Ummugul Bezirhan, Boston College
Janet Shufor Bih Epse Fofang, University of Pittsburgh
Peter Bodary, University of Michigan School of Kinesiology
Brad Bolender, Finetune by Prometric
Jill Burstein, Duolingo
Hye-Jeong Choi, HumRRO
Jinmin Chung, Univ. of Iowa
Christina Cipriano, Yale University
Lisa Clark, City University of New York
Victoria Delaney, San Diego State University
Onur Demirkaya, Riverside Insights
Scott Elliot, SEG Measurement
Andrew Emerson, National Board of Medical Examiners
Mingyu Feng, WestEd
Taiwo Feyijimi, University of Georgia
Carla Firetto, Arizona State University
Jonathan Foster, University at Albany
Samantha Goldman, The University of Kansas
Chad Green, Loudoun County Public Schools
Joe Grochowalski, College Board
Yi Gui, The University of Iowa
Aysegul Gunduz, University of Alberta
Hongwen Guo, ETS Research Institute
Yage Guo, Center for Applied Linguistics
Gulsah Gurkan, Pearson
Suhwa Han, Cambium Assessment
Michael Hardy, Stanford University
Qiwei He, Georgetown University
Alexander Hoffman, AleDev Research & Consulting
Ruikun Hou, Technical University of Munich

Ruiping Huang, University of Illinois Chicago
 Yue Huang, Measurement Incorporated
 Hiu Ching Hung, Friedrich-Alexander-Universität Erlangen-Nürnberg
 HUIMIN JIAO
 Jamie Jirout, University of Virginia
 Ji Yoon Jung, Boston College
 Olasunkanmi Kehinde, Norfolk State University
 YoungKoung Kim, The College Board
 Becky King, University of Pittsburgh
 Miryeong Koo, University of Illinois at Urbana-Champaign
 Aakash Kumar, Texas A&M University
 Alexander Kwako, Cambium Assessment
 Brandon LeBeau, WestEd
 Hansol Lee, Stanford University
 Arun Balajiee Lekshmi Narayanan, University of Pittsburgh
 Hongli Li, Georgia State University
 Tianwen Li, University of Pittsburgh
 Li Liang
 Boyuan LIU, Department of Educational Psychology, The Chinese University of Hong Kong
 Chen Liu, UC Merced
 Will Lorie
 Susan Lottridge, Cambium Assessment
 Max Lu, Harvard University
 Yi Lu, Federation of State Boards of Physical Therapy
 Wenchao Ma, University of Minnesota
 Henry Makinde, University of North Carolina - Greensboro
 Mike Maksimchuk, Kent Intermediate School District
 Salih Mansur, Touro University of New York
 Jamie Mikeska, ETS
 Mubarak Mojoyinola, University of Iowa
 Wesley Morris, Vanderbilt University
 Tim Moses, Buros Center for Testing
 William Muntean, National Council of State Boards of Nursing
 Mariel Musso, University of Granada- CONICET
 Supraja Narayanaswamy, Acelero Inc.
 Lynn Nguyen, Fritutions eTutoring
 Tram-Anh Tran Nguyen, University of Massachusetts, Amherst
 Chunling Niu, The University of the Incarnate Word
 Kai North, Cambium Learning Group, Inc.
 Teresa Ober, ETS
 Maria Oliveri, Purdue University
 Christopher Ormerod, Cambium Assessment
 Jay Parkes, University of New Mexico
 Hallie Parten, University of Virginia
 Katie Pedley, Pearson
 Benjamin Pierce, University of Pittsburgh
 Andrew Potter, Arizona State University
 Sonya Powers, Edmentum
 Ricardo Primi, Universidade São Francisco
 Sarah Quesen, WestEd
 Ruchi Sachdeva, Pearson

Fariha Hayat Salman, American University in Dubai
Lydia Scholle-Cotton, Queen's University (Kingston, ON, Canada)
Qingzhou Shi, Northwestern University
Jinnie Shin, University of Florida
Anthony Shiver, Law School Admission Council
Stephen Sireci, University of Massachusetts Amherst
Anastasia Smirnova, San Francisco State University
Xiaomei Song, Case Western Reserve University School of Medicine
Kayden Stockdale, Virginia Tech
Caitlin Tenison, ETS
Danielle Thomas, Carnegie Mellon University
Zewei Tian, University of Washington
Nhat Tran, University of Pittsburgh
FELIPE Valentini, Graduate School of Psychology, Universidade São Francisco
Marcus Walker, National Commission on Certification of Physician Assistants
Cole Walsh, Acuity Insights
Huanxiao Wang, University of Pennsylvania
Yun-Han Weng, Ohio State University
Joshua Wilson, University of Delaware
Sirui Wu, University of British Columbia
Hyesun You, University of Iowa
Meltem Yumsek Akbaba, Ministry of National Education, Turkey
Diego Zapata-Rivera, ETS
Dake Zhang, Rutgers University
Jiayi (Joyce) Zhang, University of Pennsylvania
Liang Zhang, University of Georgia
Ting Zhang, American Institutes for Research
Lauren Zito, WGU Labs

Table of Contents

<i>Input Optimization for Automated Scoring in Reading Assessment</i>	
Ji Yoon Jung, Ummugul Bezirhan and Matthias von Davier	1
<i>Implementation Considerations for Automated AI Grading of Student Work</i>	
Zewei Tian, Alex Liu, Lief Esbenshade, Shawon Sarkar, Zachary Zhang, Kevin He and Min Sun	9
<i>Compare Several Supervised Machine Learning Methods in Detecting Aberrant Response Pattern</i>	
Yi Lu, Yu Zhang and Lorin Mueller	21
<i>Leveraging multi-AI agents for a teacher co-design</i>	
Hongwen Guo, Matthew S. Johnson, Luis Saldivia, Michelle Worthington and Kadriye Ercikan	25
<i>Long context Automated Essay Scoring with Language Models</i>	
Christopher Ormerod and Gitit Kehat	35
<i>Optimizing Reliability Scoring for ILSAs</i>	
Ji Yoon Jung, Ummugul Bezirhan and Matthias von Davier	43
<i>Exploring AI-Enabled Test Practice, Affect, and Test Outcomes in Language Assessment</i>	
Jill Burstein, Ramsey Cardwell, Ping-Lin Chuang, Allison Michalowski and Steven Nydick	50
<i>Develop a Generic Essay Scorer for Practice Writing Tests of Statewide Assessments</i>	
Yi Gui	58
<i>Towards assessing persistence in reading in young learners using pedagogical agents</i>	
Caitlin Tenison, Beata Beigman Kelbanov, Noah Schroeder, Shan Zhang, Michael Suhan and Chuyang Zhang	82
<i>LLM-Based Approaches for Detecting Gaming the System in Self-Explanation</i>	
Jiayi (Joyce) Zhang, Ryan S. Baker and Bruce M. McLaren	91
<i>Evaluating the Impact of LLM-guided Reflection on Learning Outcomes with Interactive AI-Generated Educational Podcasts</i>	
Vishnu Menon, Andy Cherney, Elizabeth B. Cloude, Li Zhang and Tiffany Diem Do	99
<i>Generative AI in the K–12 Formative Assessment Process: Enhancing Feedback in the Classroom</i>	
Mike Thomas Maksimchuk, Edward Roeber and Davie Store	107
<i>Using Large Language Models to Analyze Students’ Collaborative Argumentation in Classroom Discussions</i>	
Nhat Tran, Diane Litman and Amanda Godley	111
<i>Evaluating Generative AI as a Mentor Resource: Bias and Implementation Challenges</i>	
Jimin Lee and Alena G Esposito	126
<i>AI-Based Classification of TIMSS Items for Framework Alignment</i>	
Ummugul Bezirhan and Matthias von Davier	134
<i>Towards Reliable Generation of Clinical Chart Items: A Counterfactual Reasoning Approach with Large Language Models</i>	
Jiaxuan Li, Saed Rezayi, Peter Baldwin, Polina Harik and Victoria Yaneva	142
<i>Using Whisper Embeddings for Audio-Only Latent Token Classification of Classroom Management Practices</i>	
Wesley Griffith Morris, Jessica Vitale and Isabel Arvelo	154

<i>Comparative Study of Double Scoring Design for Measuring Mathematical Quality of Instruction</i> Jonathan Kyle Foster, James Drimalla and Nursultan Japashov	163
<i>Toward Automated Evaluation of AI-Generated Item Drafts in Clinical Assessment</i> Tazin Afrin, Le An Ha, Victoria Yaneva, Keelan Evanini, Steven Go, Kristine DeRuchie and Michael Heilig	172
<i>Numeric Information in Elementary School Texts Generated by LLMs vs Human Experts</i> Anastasia Smirnova, Erin S. Lee and Shiyang Li	183
<i>Towards evaluating teacher discourse without task-specific fine-tuning data</i> Beata Beigman Klebanov, Michael Suhan and Jamie N. Mikeska	192
<i>Linguistic proficiency of humans and LLMs in Japanese: Effects of task demands and content</i> May Lynn Reese and Anastasia Smirnova	201
<i>Generative AI Teaching Simulations as Formative Assessment Tools within Preservice Teacher Preparation</i> Jamie N. Mikeska, Aakanksha Bhatia, Shreyashi Halder, Tricia Maxwell, Beata Beigman Klebanov, Benny Longwill, Kashish Behl and Calli Shekell	212
<i>Using LLMs to identify features of personal and professional skills in an open-response situational judgment test</i> Cole Walsh, Rodica Ivan, Muhammad Zafar Iqbal and Colleen Robb	221
<i>Automated Evaluation of Standardized Patients with LLMs</i> Andrew Emerson, Le An Ha, Keelan Evanini, Su Somay, Kevin Frome, Polina Harik and Victoria Yaneva	231
<i>LLM-Human Alignment in Evaluating Teacher Questioning Practices: Beyond Ratings to Explanation</i> Ruikun Hou, Tim Fütterer, Babette Bühler, Patrick Schreyer, Peter Gerjets, Ulrich Trautwein and Enkelejda Kasneci	239
<i>Leveraging Fine-tuned Large Language Models in Item Parameter Prediction</i> Suhwa Han, Frank Rijmen, Allison Ames Boykin and Susan Lottridge	250
<i>How Model Size, Temperature, and Prompt Style Affect LLM-Human Assessment Score Alignment</i> Julie Jung, Max Lu, Sina Chole Benker and Dogus Darici	265
<i>Assessing AI skills: A washback point of view</i> Meirav Arieli-Attali, Beata Beigman Klebanov, Tenaha O'Reilly, Diego Zapata-Rivera, Tami Sabag-Shushan and Iman Awadie	274
<i>Using Generative AI to Develop a Common Metric in Item Response Theory</i> Peter Baldwin	281
<i>Augmented Measurement Framework for Dynamic Validity and Reciprocal Human-AI Collaboration in Assessment</i> Taiwo Feyijimi, Daniel O Oyeniran, Oukayode Apata, Henry Sanmi Makinde, Hope Oluwaseun Adegoke, John Ajamobe and Justice Dadzie	290
<i>Patterns of Inquiry, Scaffolding, and Interaction Profiles in Learner-AI Collaborative Math Problem-Solving</i> Zilong Pan, Shen Ba, Zilu Jiang and Chenglu Li	297
<i>Pre-trained Transformer Models for Standard-to-Standard Alignment Study</i> Hye-Jeong Choi, Reese Butterfuss and Meng Fan	306

<i>From Entropy to Generalizability: Strengthening Automated Essay Scoring Reliability and Sustainability</i>	Yi Gui	312
<i>Undergraduate Students' Appraisals and Rationales of AI Fairness in Higher Education</i>	Victoria Delaney, Sunday Stein, Lily Sawi and Katya Hernandez Holliday	329
<i>AI-Generated Formative Practice and Feedback: Performance Benchmarks and Applications in Higher Education</i>	Rachel van Campenhout, michelle weaver clark, Jeffrey S. Dittel, Bill Jerome, Nick Brown and Benny Johnson.....	337
<i>Beyond Agreement: Rethinking Ground Truth in Educational AI Annotation</i>	Danielle R Thomas, Conrad Borchers and Ken Koedinger	345
<i>Automated search algorithm for optimal generalized linear mixed models (GLMMs)</i>	Miryeong Koo and Jinming Zhang	352
<i>Exploring the Psychometric Validity of AI-Generated Student Responses: A Study on Virtual Personas' Learning Motivation</i>	Huanxiao Wang.....	359
<i>Measuring Teaching with LLMs</i>	Michael Hardy	367
<i>Simulating Rating Scale Responses with LLMs for Early-Stage Item Evaluation</i>	Onur Demirkaya, Hsin-Ro Wei and Evelyn Johnson	385
<i>Bias and Reliability in AI Safety Assessment: Multi-Facet Rasch Analysis of Human Moderators</i>	Chunling Niu, Kelly Bradley, Biao Ma, Brian Waltman, Loren Cossette and Rui Jin	393
<i>Dynamic Bayesian Item Response Model with Decomposition (D-BIRD): Modeling Cohort and Individual Learning Over Time</i>	Hansol Lee, Jason B. Cho, David S. Matteson and Benjamin Domingue	398
<i>Enhancing Essay Scoring with GPT-2 Using Back Translation Techniques</i>	Aysegul Gunduz, Mark Gierl and Okan Bulut	406
<i>Mathematical Computation and Reasoning Errors by Large Language Models</i>	Liang Zhang and Edith Graf.....	417