

AIME-Con 2026

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

Volume 2: Works in Progress

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 979-8-9983004-1-7

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 Work-in-Progress Papers

AIME-Con invited submissions of Working Papers, which were submissions of up to 1,000 words in the form of a structured summary. This format was designed for work-in-progress or pilot studies. Working Papers required a title, short abstract, and followed a structured format with the following headings:

- Background
- Aims
- Sample(s)
- Methods
- Results (or Anticipated Results)
- Conclusions (or Anticipated Contributions)
- References, tables, and figures included as needed

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 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

Ketan , University of Massachusetts Amherst
Tazin Afrin, NBME
Norah Almousa, University of Pittsburgh
Talal Alzabidi, University of North Carolina Greensboro
Kylie Anglin, University of Connecticut
Olukayode Apata, Texas A&M University
Sergio Araneda, Caveon
Burcu Arslan, ETS Research Institute
Aydan Azimzade, Teachers College, Columbia University
Damilola Babalola
Justin Barber, Pearson
Mark Beck, Measurement Incorporated
Ahmed Bediwy, Pearson
Kashish Behl, ETS
Magdalen Beiting-Parrish, EdAlfy, CUNY Graduate Center
Ummugul Bezirhan, Boston College
Obed Boateng, Virginia Commonwealth University
Kelly Boden, Amira Learning
Brad Bolender, Finetune by Prometric
Laine Bradshaw, Pearson
Jill Burstein, Duolingo
Leon Camus, DIPF | Leibniz Institute for Research and Information in Education
Roti Chakraborty, American Institutes for Research
Shubham Chakraborty, Arizona State University
Jie Chen, Measurement Incorporated
Siyuan Chen, Duolingo Inc.
Ikkyu Choi, ETS
Jeongwon Choi, Vanderbilt University
S. Thomas Christie, Renaissance Philanthropy
Michael Chrzan, Center for Educational Data Science and Innovation, University of Maryland
Ruhan Circi, American Institutes for Research
Dante Cisterna, ATLAS - University of Kansas
Amy Clark, ATLAS: University of Kansas
Susan Davis-Becker, ACS Ventures LLC
Victoria Delaney, San Diego State University
Onur Demirkaya, Riverside Insights
Rahul Divekar, Bentley University
Svetlana Dmitrieva, University of South Carolina
Enis Dogan, NCES

Andrew Emerson, National Board of Medical Examiners
Farina Faiz, George Mason University
Fen Fan, National Board of Medical Examiners
Meng Fan, HUMRRO
Michael Fauss, ETS
Yetunde Fawehinmi, Texas A&M University
Peter Foltz, University of Colorado - Boulder
Julius Frost
Jie Gao, McGill University
Guher Gorgun, University of Georgia
Irina Grabovsky
Edith Graf, Proof-of-Concept Learning LLC
Joe Grochowalski, College Board
Yi Gui, Measurement Inc
Suhwa Han, Cambium Assessment
Polina Harik, NBME
Brian Harrold, University of South Carolina
Yong He
Michael Hemenway, Pearson
Langdon Holmes
Shahin Hossain
Jing Huang, Purdue University
Sijia Huang, Indiana University Bloomington
Wei Huang, University of Alabama, College of Education
Yue Huang, Measurement Incorporated
Fareya Ikram, University of Massachusetts Amherst
Megan Imundo, WGU Labs
Gregory Jacobs, Pearson
Paul Jewsbury, Duolingo
Yanlin Jiang, NCCPA
Ji Yoon Jung, Boston College
Julie Jung, Harvard University
Tzur Karelitz, National Institute for Testing & Evaluation (NITE)
Farzan KarimiMalekabadi, Edmentum
Euigyum Kim, Sogang University
Eileen Koh, Purdue University
Harvey Ngole Kolle, University of Alberta
Miryeong Koo, University of Illinois at Urbana-Champaign
Tianqi Lang
Antonio Laverghetta Jr., Wesleyan University
Thuy Dieu Le, University of Massachusetts Amherst
Sophia Lebaschi, California State University Northridge
Haneul Lee, University of Massachusetts Amherst
Jimin Lee, Clark University
Will Lee, UMass Amherst
Youlim Lee, University of Massachusetts Lowell
Arun Balajiee Lekshmi Narayanan, University of Pittsburgh
Jennifer Lewis, University of Massachusetts Amherst
Jujia Li, University of Alabama
Tianwen Li, University of Pittsburgh
Zhifei Li

Guangming Ling, ETS
 Helena Liu, West Lafayette Junior/Senior High School
 Xiang Liu, ETS
 Yong Luo, Pearson
 Meng Lyu
 Jing Ma, University of Iowa
 Ye Ma, Amazon Web Services
 Hotaka Maeda, Smarter Balanced
 Henry Makinde
 Michael McCarthy, Education Collaboratory at Yale University
 Vandhana Mehta, Pearson
 Steven Moore, George Mason University
 Josh Moskowitz, Acuity Insights
 William Muntean, National Council of State Boards of Nursing
 NaRa Nam
 Sungjin Nam, ACT Education Corp.
 Anna Nasyrova, University of Massachusetts Amherst
 Chunling Niu, The University of Texas at San Antonio
 Teresa Ober, ETS
 Burhan Ogut, American Institutes for Research
 Christopher Ormerod, College Board
 Corey Palermo, Measurement Incorporated
 Jung Yeon Park, George Mason University
 Hallie Parten, University of Virginia
 Benjamin Pierce, Institute for Intelligent Systems
 Prashreet Poudel, Western Carolina University
 Jiten Pradhan, American College of Cardiology
 Emily Proctor, College of Engineering, Purdue University
 Aleena Raj, Kaplan North America
 Ranjidha Rajan, MSU Denver
 Brandi Robinson
 Alex Rudniy, Drew University
 Christopher Runyon, National Board of Medical Examiners
 Ruchi Sachdeva, Pearson
 Tahira Saddiqa, university of North Texas
 Zachary Schultz, Cambium Learning Group Inc.
 Jonathan Schweig, RAND
 Qian Shen, University of Florida
 David Shin, American Board of Psychiatry and Neurology
 Jinnie Shin, University of Florida
 Anthony Shiver, Law School Admission Council
 Gozde Sirganci, Southern Methodist University
 William Skorupski, Amira Learning
 Dan Song, The University of Iowa
 Angela Starrett, University of South Carolina
 Gordon Stein, Renaissance Philanthropy
 Sunday Stein, San Diego State University & University of California
 Javier Suárez-Álvarez, University of Massachusetts Amherst
 Susha Suresh, Acuity Insights
 Judy Tang, Westat
 Steven Tang, eMetric LLC

Ethan Tenison, Amira Learning
 Danielle Thomas, Carnegie Mellon University
 Yu Tian, Arizona State University
 Lingbo Tong, University of Wisconsin–Madison
 Kazi Md Rukan Uddin, University of Houston
 Kirk Vanacore, Cornell University
 Claudia Ventura, University of Connecticut
 Yen Vo, The University of Iowa
 Matthias von Davier, Boston College
 Mai Vuong, University of Pennsylvania
 Cole Walsh, Acuity Insights
 Kuo Wang, Southern Methodist University
 Ling Wang
 Ting Wang
 Zuowei Wang, Educational Testing Service
 Field Watts, ETS
 Yumou Wei, Carnegie Mellon University
 David Whitcomb, ATLAS, University of Kansas
 John Whitmer, Learning Data Insights
 Joshua Wilson, University of Delaware
 John Workman, EQAO
 Lixin Wu, University of Illinois at Urbana-Champaign
 Sirui Wu, University of British Columbia
 Yi-Chen Wu, National Center on Educational Outcomes, University of Minnesota
 Yi-Fang Wu, Cambium Assessment Inc.
 Michael Xiao, University of Washington
 Weixuan Xiao, Teachers College, Columbia University
 Dawei Xie, Northwestern University
 Tianyu Xu, Teachers College, Columbia University
 Yue Xu, University of Florida
 Kang Xue, Cambium Assessment
 Rohit Yadav, University of Kansas
 Daijin Yang, Northeastern University
 Yaxuan Yang, University of Georgia
 Yiting Yao, Florida State University
 Farzana Yasmin, University of Alberta
 Michelle Yin, Northwestern University
 Jiseung Yoo
 Jessica Younger, Digital Promise
 Kefan Yu
 Idris Yusuf, University of Houston
 Hongyi Zhai, North Oconee High School
 Liang Zhang, University of Michigan
 Qiongyao Zhang, Robert Morris University
 Shan Zhang, University of Florida
 Yuxiao Zhang, Purdue University
 Zhehan Zhang, Clemson University

Table of Contents

<i>A Cognitive Lab Study of Student AI Chatbot Use in Online Coursework</i> Jinah Choi, Sonya Powers, Farzan Karimi-Malekabadi and Michelle Barrett	1
<i>A Preliminary Semantic Screen for IRT Local Dependence</i> Qian Shen	9
<i>Evaluating Multi-Phase and Granular Strategies for Evidence-Anchored Automated Scoring</i> Alessia Marigo, Laura Wright, Linda Malkin and Xin Xie	17
<i>Consensus without Accuracy: Investigating LLM’s Recovery of Item Difficulty Using Paired Comparisons</i> Michael Leon Chrzan and Benjamin Domingue	24
<i>Implementing Generative AI Assessment Guidelines in a Mexican University</i> Melchor Sánchez-Mendiola, Enrique Buzo-Casanova, Elibidú Ortega-Sánchez, Manuel García-Minjares and Adriana Cruz Juárez	37
<i>TRACE: Automated Multi-Granularity Analysis of Text Revision</i> Yu Tian, Katerina Christhlf, Andrew Potter, Jessica Early, Steve Graham and Danielle McNamara	45
<i>Comparing θ Representation Strategies for Reconstructing Item Characteristic Curves</i> Sungjin Nam	52
<i>Assessing Large Language Model Performance in Post-Certification-Examination Comment Categorization</i> Huaping Sun, Kristin O’Brien, Colleen Burke Kave, David Shin, Qiao Lin and Jeffrey Marc Lyness	57
<i>Rethinking Pilot Data: Evaluating LLM Synthetic Data for Scale Development</i> Margarita Olivera-Aguilar, Samuel H. Rikoon, Paul D. Bailey, Michael B. Kruse and Kamal Middlebrook	62
<i>Hierarchical Attention Network Architecture for Automated Response Time Prediction</i> Wei Shuang Schneider	69
<i>Beyond Agreement: Calibrating Automated Scoring to Human Judgment Using Control Scripts</i> Mark Dulhunty, Alex Codoreanu and Nathan Zoanetti	79
<i>Human-in-the-Loop Gemini Item Generation for Large-Scale MCQ Banks</i> Melchor Sánchez-Mendiola, Milton García-Lima, Elibidú Ortega-Sánchez, Manuel García-Minjares and Enrique Buzo-Casanova	86
<i>Score Comparability Limits with Common AES Validation Measures</i> Tim Moses, Seongeun Kim and Nick Trout	93
<i>Finding Evidence of Communication Behaviors using LLMs</i> Tazin Afrin, Saed Rezayi, Janet Mee, Polina Harik and Le An Ha	100
<i>Designing and Evaluating a Measurement-Oriented RAG Framework for ELL Writing Assessment</i> Daeryong Seo	104
<i>Automated Item Evaluation: Predicting Item Acceptance and Rejection using LLM-Generated Critiques</i> Hotaka Maeda and Yikai Lu	113

<i>Production-Ready Automated Item Generation in Educational Assessment: Integration with Operational Workflows</i>	
Hotaka Maeda, Kargi Chauhan and Tharunya Chandrashekar	130
<i>Using Generative AI to Simulate Item Responses by Skills Insight Score Band</i>	
Jianshen Chen, Chansoon Lee, Kylie Gorney and Judit Antal	140
<i>Exploring AI-driven Methods for Pre-calibrating Difficulty of Mathematics Items with Visual Content</i>	
Leah Walker and Jinnie Choi	145
<i>Evaluating Multiple Models for Predicting Item Difficulty in a Principled Assessment Design</i>	
Alexandra Lane Perez, Christina Schneider, Sangdon Lim, Garron Gianopulos and Kang Xue	152
<i>Transformer-Aided Detection of Gaming in Constructed-Response English Language Assessments</i>	
Melanie Sharif, Scott Hellman, Martha Bellows and Sue Lottridge	159
<i>Calibrated, Interpretable Automated Scoring with LLM Likelihoods</i>	
Thomas Christie, Markus Hauru and Anna Rafferty	168
<i>Predicting Item-to-Range Performance Level Descriptor Matches with Structured Language Model Reasoning</i>	
Kang Xue, M. Christina Schneider, Sangdon Lim, Garron Gianopulos and Alexandra Perez .	177
<i>Using LLM Judges' Paired Comparisons to Estimate Mathematics Item Difficulty</i>	
Lanrong Li, Mohammad A. A. Abulela and Matthew Gushta	184
<i>Math-Specialized or General-Purpose? A Comparison of MathBERT and DeBERTa-v3-large for Automated Scoring of Mathematics-Explanation Items</i>	
Gregory M. Jacobs, Ahmed H. Bediwy and Martha Bellows	193
<i>How Much Does Hyperparameter Tuning Actually Help? An Efficiency Survey for Fine-Tuning Transformers to Score Mathematics-Explanation Items</i>	
Gregory M. Jacobs, Ahmed H. Bediwy and Martha Bellows	200
<i>Responsible use of generative AI when creating reading comprehension questions: Inference matters</i>	
Zuowei Wang, Michael Flor, Jiyun Zu, Tenaha O'Reilly and Wanjing Anya Ma	207
<i>Fine-tuning DeBERTa-v3 to Automate Spatial Language Classification</i>	
Sam Agnoli, Qingzhou Shi and David Uttal	211
<i>CDM Q-Matrix Discovery with LLMs: Fusing Domain Knowledge with Empirical Evidence</i>	
Susu Zhang and V. N. Vimal Rao	216
<i>Composite Scores vs. Preference Rankings: Measuring Architecture Effects in LLM Feedback</i>	
Harvey Ngoe Kolle, Carrie Demmans Epp, Amna Liaqat and Maria Cutumisu	231
<i>Identifying Communication Profiles from Physician–Patient Conversations Using Large Language Models</i>	
Reyhaneh Hosseinpourkhoshkbari, Yash Bipin Jain and Richard Golden	235
<i>LLM-Based Pairwise Judgment for Math Item Parameter Modeling Using Workflows and Agents</i>	
Suhwa Han and Frank Rijmen	248
<i>Beyond Volume: How Cognitive Fingerprints Predict Residual Gain in AI Tutoring</i>	
Aleena K Raj and Nina Deng	260

<i>Automatic Prompt Engineering for Generative AI-Based Essay Scoring</i> Yue Huang	267
<i>A Comprehensive Evaluation of GenAI-based Items in a Medical Examination</i> Yanlin Jiang, Marcus Walker, Andrew Dallas, Aquia Richburg, Nikole Gregg and Brittany Corrigan	275
<i>When One Benchmark Hides Many Truths: Mixture-IRT and LLM Difficulty Prediction</i> Enis Dogan and Burhan Ogut	282
<i>LLM-As-An-Assessor: Can Open-Weight LLMs Assess Computational Thinking via Student-Designed Embodied Games?</i> William Lee, Sai Gattupalli, Ivon Arroyo and Brendan O'Connor	289
<i>Funnel Plot Analysis of Examinee Performance in Longitudinal Assessment</i> Aquia Richburg and Marcus Walker	298
<i>How Much Training Data Is Enough? Fine-Tuning LLMs for Short-Answer Scoring</i> Joshua A McGrane and David Torres Irribarra	306
<i>Evaluating Machine Learning, Transformer, and LLM Approaches for Large-Scale Automated Essay Scoring</i> Cyrus Sai-Cheong Chan, Charles Anifowose, Su Zhang, Winnie Wing-Yee Tse, Nizam Radwan and Hyunah Kim	315
<i>Construct Validity of Small-Sample Transformer Scoring Models: A Mechanistic Interpretability Approach</i> Michael P. Hemenway and Martha Bellows	322
<i>Applying Evidence-Centered Design to Automated Evals of AI-Powered Assessment Systems</i> Kristen DiCerbo, Britte Haugan Cheng and John Whitmer	329
<i>MIST: Misconception Identification and Sycophancy Tracking</i> Lionel Meng, Lingbo Tong and Icy Zhang	335
<i>Building Teacher Capacity to Generate and Evaluate Math Items with Small LLMs</i> Minseok Kim	344
<i>Principled Approaches to Building AI Partners for Assessing and Supporting Small-Group Collaboration</i> Peter W Foltz, Chelsea Chandler and Mon-Lin Monica Ko	350
<i>Measuring Writing Proficiency Through Keystroke Process Data: A Cross-School Generalizability Study</i> Damilola Babalola, Naga Buddarapu, Caleb Scott, Piotr Mitros, Paul Deane and Collin Lynch	358
<i>Self-Revising Agents as Item Writers: Benchmarking Against Professional Item Writing</i> Steven Tang, Zhen Li and Richard Patz	369
<i>Evaluating Neural Network-Based IRT in Multistage Adaptive Testing with Small Item Banks</i> Seong Eun Hong	377
<i>Validating LLM-Rated Item Features for Explanatory Item Response Models</i> Ahmed H. Bediwy and Mubarak O. Mojoyinola	382
<i>Efficacy of Student-AI Co-Authored Math Word Problems in an Intelligent Tutoring System</i> Kole Norberg, April Murphy, Steve Fancsali and Steve Ritter	389

<i>Assessing Small Language Models as Decimal-Arithmetic Tutors: A Measurement Framework</i>	
Mai Que Vuong, Shahana Ahmadli, Michelle Zhou, Minseok Kim, Shruti Mehta, Talita de Paula Cypriano de Souza and Seiji Isotani	397
<i>LLM as Investigator and Judge in Pairwise Comparisons for Item Parameter Modeling</i>	
Brian Harrold, Vincent Fakiyesi, Maria D'Brot, Susan Lottridge, Justin Barber and Michael Hemenway	405
<i>LLM-Based Metadiscourse Analysis of Demographic Variations in ELL Hedge Use</i>	
Xinyu Xu and Tianyu Xu	416
<i>Beyond the Grid: Auditing Code-Editing Search for Automated Essay Scoring</i>	
Julius Frost	423
<i>LLM Difficulty Prediction Across Three L2 Task Types and Two Examination Forms</i>	
Meng Lyu, Moses Oluoke Omopekunola and Feng Wang	441
<i>Measuring Science Teachers' Generative AI Literacy for Classroom Assessment: Performance Assessment</i>	
Ruiping Huang and Yue Yin	448
<i>Do Synthetic Students Thrive? Validity and Fairness of Generated Assessment Data</i>	
Neba Nfonsang and Temple Lovelace	459
<i>AI-Enabled Quality Assurance for Multiple-Choice Assessment Items</i>	
Steven James Moore and Nicholas Diana	464
<i>More than Words: Modeling What Teachers Praise and How They Praise It</i>	
Jiseung Yoo, Paiheng Xu and Jing Liu	472
<i>Does an Automated Source Integration Score Respond to Revision? Extending Evidence of Construct Validity with a Feedback Experiment</i>	
Andrew Potter, Yu Tian, Kaitlin Van Houghton, Manmeet Singh, Renu Balyan, Maria Goldshtein, Laura K. Allen and Danielle S. McNamara	483