

AIME-Con 2026

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

**Volume 1: Full 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](https://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](https://PearsonAssessments.com/navvy) to learn more**

800-627-7271 | [PearsonAssessments.com](https://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](https://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](https://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](mailto:ncme@ncme.org)

ISBN 979-8-9983004-0-0

## 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](http://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 ( $\leq 12$  words), a brief abstract ( $\leq 50$  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 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 Kelle, 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>Stochastic Constrained Test Assembly for AI-Enabled Assessment Systems</i><br>Alina A. von Davier .....                                                                                                                        | 1   |
| <i>Psychmet- Measurement Foundational Competencies ChatBot</i><br>Henry Makinde, Hope Adegoke and Mubarak Mojoyinola .....                                                                                                        | 7   |
| <i>Exploratory Bayesian Multidimensional IRT with Rater Effects for Scoring Generative AI Usage</i><br>Leon Camus, Sebastian Gombert, Fabíola Ribeiro, Gianluca Romano, Carmen Köhler and Hendrik Drachsler .....                 | 13  |
| <i>Predicting IRT Parameters for Passage-Based Reading Items Using NLP</i><br>Ruitao Liu .....                                                                                                                                    | 30  |
| <i>Prior-Informed 3PL Calibration: Reducing Sample Size via Predictive Modeling</i><br>Ruitao Liu and Aixin Tan .....                                                                                                             | 36  |
| <i>Connecting Student Feedback Understanding to Gains in Argumentative Writing and Revision</i><br>Diane Litman, Zhexiong Liu, Tianwen Li, Lindsay Clare Matsumura, Richard Correnti and Elaine Wang .....                        | 43  |
| <i>Evaluating Feedback Focus and Pedagogical Adaptivity in LLM Feedback on Student Writing</i><br>Norah Almousa, Shayan Peyghambari Oskoui, Raquel Coelho, Gayle Rogers, Xiang Lorraine Li and Diane Litman .....                 | 55  |
| <i>S2A3: Thompson Sampling and Stochastic Exposure Control for High-Stakes CATs</i><br>James Sharpnack, Alexander Tsigler, J.R. Lockwood, Steven Nydick and Alina A. von Davier .....                                             | 76  |
| <i>Multimodal Item Parameter Estimation using Simulated Response Probabilities</i><br>Christopher Michael Ormerod and YoungKoung Kim .....                                                                                        | 83  |
| <i>Can LLMs Replace Semantic Similarity for Scoring Student Code Explanations?</i><br>Arun Balajee Lekshmi Narayanan, Mohammad Hassany and Peter Brusilovsky .....                                                                | 92  |
| <i>Developing and Validating an Automatic Scoring Model for Chatbot-Based Conversational Speech</i><br>Danwei Cai, Audrey Kittredge, Ben Naismith, Xiangying Jiang and Kevin Yancey .....                                         | 100 |
| <i>Fractional Information Gain: A Personalization Metric for Black-Box Predictive Models</i><br>S. Thomas Christie, Markus Hauru and Anna N. Rafferty .....                                                                       | 109 |
| <i>Autoscoring Anticlimax: A Meta-analytic Understanding of AI's Short-answer Shortcomings and Word-ing Weaknesses</i><br>Michael Hardy .....                                                                                     | 117 |
| <i>A Multi-Agent Architecture for Valid, Reliable, and Scalable Skills Assessment</i><br>Megan N Imundo, Kjorte Harra, Lesley Reilly, Cory Hammon, Lauren Zito, Catrina Nieser and Betheny Gross .....                            | 139 |
| <i>Speed–Accuracy Trade-offs in Knowledge Component Identification from Student Code</i><br>Jing Fan, Tsvetomila Mihaylova, Peter Brusilovsky, Juho Leinonen, Charles Koutcheme, Narges Norouzi, Bitá Akram and Arto Hellas ..... | 146 |
| <i>Effects of Demographic Representations and Model Training Methods on Automated Scoring Engines</i><br>Yangmeng Xu, Martha Bellows and Edward W Wolfe .....                                                                     | 154 |

|                                                                                                                                                                               |     |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| <i>Design Choices in Encoder-Based IRT-3PL Parameter Prediction for Passage-Based Multiple-Choice Questions</i>                                                               |     |
| Sungjin Nam .....                                                                                                                                                             | 161 |
| <i>It Quacks Like a Response: Evaluating Synthetically-Generated Responses in Assessment Research</i>                                                                         |     |
| Cole Walsh and Rodica Ivan .....                                                                                                                                              | 170 |
| <i>Automated Justification-Depth Scoring for Adaptive Support in AI-Supported Instructional Tasks</i>                                                                         |     |
| Songhee Han, Jueun Shin and Jiyeon Han .....                                                                                                                                  | 179 |
| <i>Using Natural Language Processing to Explore Alignment Between Skill Taxonomies</i>                                                                                        |     |
| Teresa M. Ober and Michael Flor .....                                                                                                                                         | 187 |
| <i>A Feasibility Study on Retrieval Practice Using Large Language Models</i>                                                                                                  |     |
| Marcus Leong, Jennifer Rose, Lisa Dierker and Antonio Laverghetta Jr. ....                                                                                                    | 194 |
| <i>Exploring Asymmetric Academic Integrity Signals in Search- and LLM-Assisted Learning</i>                                                                                   |     |
| Rahul R. Divekar .....                                                                                                                                                        | 204 |
| <i>An LLM-Enhanced Score Reporting Assistant for Teachers: System Design and Usability Evaluation</i>                                                                         |     |
| Shan Zhang, Caitlin Tenison, Diego Zapata-Rivera, Reginald Gooch and Maya Israel .....                                                                                        | 212 |
| <i>Examining LLM-Surprisal as an Indicator of Naturalness for Japanese Automated Essay Scoring</i>                                                                            |     |
| Akari Osumi, Jingying Hu, Yan Cong and Atsushi Fukada .....                                                                                                                   | 221 |
| <i>A Validity Threat Framework for Measuring Student Generative AI Use</i>                                                                                                    |     |
| Olukayode Emmanuel Apata, Yetunde Omoyiwola Fawehinmi, Daniel Olutola Oyeniran, Glory Onize Saidu, Segun Timothy Ajose, Naphtali Onalo and Oluwasegun Matthew Amoniyani ..... | 230 |
| <i>Using XGBoost to Construct a Vertically Skill Difficulty Scale</i>                                                                                                         |     |
| John Bielinski .....                                                                                                                                                          | 239 |
| <i>Evaluation of AI-generated Passages for Reading Practice: An application of causal forests</i>                                                                             |     |
| Walter L Leite, Qian Shen, Wonchae Lee, Sierra Evans, Sienna Perez, Wei Li and Jinnie Shin                                                                                    | 245 |
| <i>Supporting Distractor Quality Review Through Interpretable Semantic and Lexical Diagnostics</i>                                                                            |     |
| Michelle Banawan, Shubham Chakraborty, Katerina Christliff, Linh Huynh, Tracy Arner and Danielle McNamara .....                                                               | 253 |
| <i>Comprehensive Pipeline for Multilingual GenAI Scoring</i>                                                                                                                  |     |
| Ji Yoon Jung, Ummugul Bezirhan and Matthias von Davier .....                                                                                                                  | 260 |
| <i>From Revision to Transfer: Longitudinal Effects of LLM Writing Feedback</i>                                                                                                |     |
| Yigal Attali, Bryan Smith, Kai-Ling Lo, Andrew Runge and Jacqueline Church .....                                                                                              | 267 |
| <i>An Evaluation of Agreement and Uncertainty in LLM-Based Automated Essay Scoring</i>                                                                                        |     |
| Yiting Yao, Yanyun Yang and Huan (Hailey) Kuang .....                                                                                                                         | 280 |
| <i>Reliability Estimation Methods for Bayesian Knowledge Tracing Mastery Classifications</i>                                                                                  |     |
| Samuel D. Ihlenfeldt and Stephen E. Fancsali .....                                                                                                                            | 288 |
| <i>Automated Approaches for Scoring Math Misunderstandings in Student Self-Explanations</i>                                                                                   |     |
| Scott Crossley, Bethany Rittle-Johnson, Rebecca Adler, L Burleigh, Jules King and Meg Benner                                                                                  | 297 |
| <i>Interpretable Humans, Alien LLMs: Expert Analysis of Latent Structures in Assessment Responses</i>                                                                         |     |
| Alona Strugatski, Licol Zeinfeld, Jason Cooper, Shelley Rap, Gil Schwartz and Giora Alexandron                                                                                | 303 |

|                                                                                                                  |     |
|------------------------------------------------------------------------------------------------------------------|-----|
| <i>Bayesian Consensus Calibration of Continuously Evolving IRT Item Banks</i>                                    |     |
| Paul A Jewsbury, Steven W Nydick, Manqian Liao and Siyuan (Marco) Chen .....                                     | 314 |
| <i>On Promoting Individual-Level Fairness in Automated Scoring</i>                                               |     |
| Michael Fauss, Matthew S. Johnson and Ikkyu Choi .....                                                           | 323 |
| <i>CAP: Cross-Ability Preferences to Align Student Simulators for Item Difficulty Prediction</i>                 |     |
| Nigel Fernandez, Alexander Scarlatos, Christopher Ormerod and Andrew Lan .....                                   | 332 |
| <i>Auxiliary Information for Semantic Clustering of Assessment Items</i>                                         |     |
| Josiah Hunsberger, Aquia Richburg and Marcus Walker .....                                                        | 345 |
| <i>Assessment Use Guide Builder: Leveraging Use Cases to Generate Interpretive Assessment Materials</i>          |     |
| Yaning Cao and Nathan Dadey .....                                                                                | 355 |
| <i>Anchored Bradley-Terry Calibration Using LLM Comparative Judgments</i>                                        |     |
| Ummugul Bezirhan and Matthias von Davier .....                                                                   | 365 |
| <i>Seven Ways to Cut a Score: Mapping Essay-Scoring Predictions to Ordinal Ratings</i>                           |     |
| Ahmed H. Bediwy, Martha Bellows and Sue Lottridge .....                                                          | 372 |
| <i>Math Item Difficulty Prediction with Multimodal Input</i>                                                     |     |
| Kai North, Christopher Ormerod, Alexander Kwako and Suhwa Han .....                                              | 380 |
| <i>Generative Language Models for Argumentation Annotation</i>                                                   |     |
| Kai North and Christopher Ormerod .....                                                                          | 387 |
| <i>Prompting as Measurement: Enhancing Coding for Systematic Reviews with Large Language Models</i>              |     |
| Dandan Chen Kaptur, Yue Huang, Yanhui Guo and Xuejun Ryan Ji .....                                               | 398 |
| <i>Measuring the User, Not the Model: Critical AI Literacy</i>                                                   |     |
| Derek Briggs .....                                                                                               | 407 |
| <i>Long-Context Transformer and State-Space Architectures for Data-Efficient Automated Essay Scoring</i>         |     |
| Justin O Barber, Michael P Hemenway, Martha Bellows and Susan Lottridge .....                                    | 413 |
| <i>Can LLMs Interpret Psychometric Parameters? Quantifying "Over-Knowledge" in LLM Student Simulation</i>        |     |
| Kazunori Fukuhara and Benjamin Domingue .....                                                                    | 421 |
| <i>Fine-Tuning Large Language Models for Codebook-Guided Coding of Students' Mathematics Metaphor Responses</i>  |     |
| Liang Zhang, Stephen Hwang, Yue Ma and Jinfa Cai .....                                                           | 433 |
| <i>Lost Without Translation? Multilingual Sentence Embeddings for Linguistic-integrated Reliability Auditing</i> |     |
| Ummugul Bezirhan, Ji Yoon Jung and Matthias von Davier .....                                                     | 443 |
| <i>To Flag or Not to Flag? Detecting Concerning Content in Situational Judgment</i>                              |     |
| Susha Suresh, Cole Walsh, Rodica Ivan and Colleen Robb .....                                                     | 450 |
| <i>Two Uses of Human-scored Anchors: Few-shot Direct Scoring and Anchor-based Comparative Grading</i>            |     |
| Prashreet Poudel, Huayue Gu, Collin Lynch and Zhikai Gao .....                                                   | 458 |
| <i>Semantic Similarity is Not Enough for Comparing LLM and Human Rater Rationales</i>                            |     |
| Julie Jongeun Jung, Max Lu, Dogus Darici and Emilia Brüegge .....                                                | 466 |

|                                                                                                                                                                |     |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| <i>Automated Scoring of Oral Reading Fluency: An examination of validity evidence</i>                                                                          |     |
| Walter L Leite, Krishna Sudeep Kumar, Jaiden Magnan, Stephanie Hammerschmidt-Snidarich, Seyedahmad Rahimi and Zoey Liu .....                                   | 472 |
| <i>Image-Based Representation of Item Response Patterns for Test Integrity</i>                                                                                 |     |
| Akshay Badola and Mo Zhang .....                                                                                                                               | 483 |
| <i>Fine-tuning Large Language Models for Automated Scoring: Classification, Regression, vs. Ordinal Regression</i>                                             |     |
| Ji Yun Zu and Akshay Badola .....                                                                                                                              | 490 |
| <i>Simulating Examinee Responses to Figure Matrices Items Using VLMs for Item Evaluation</i>                                                                   |     |
| Onur Demirkaya and Evelyn Johnson .....                                                                                                                        | 497 |
| <i>A Validity Argument Framework for Automated Item Generation Systems</i>                                                                                     |     |
| Euigyum Kim, Hyo Jeong Shin and Alina von Davier .....                                                                                                         | 506 |
| <i>Hierarchical Analytic Writing Feedback with Fine-Tuned Transformers</i>                                                                                     |     |
| Michael P. Hemenway, Justin O. Barber, Martha Bellows and Susan Lottridge .....                                                                                | 516 |
| <i>Source-Text-Conditioned Cloze Generation for Discourse Comprehension Assessment in Intelligent Textbooks</i>                                                |     |
| Langdon Holmes, Wesley Morris, Scott Andrew Crossley and Aiden Min .....                                                                                       | 525 |
| <i>Assessing the Reliability and Construct Representation of LLM-based Language Proficiency Scores</i>                                                         |     |
| Langdon Holmes, Scott Andrew Crossley, Joon Suh Choi and Wesley Morris .....                                                                                   | 534 |
| <i>A Multi-Year Investigation of Undergraduates' Generative AI Fairness Appraisals in Higher Education</i>                                                     |     |
| Sunday Stein, Victoria Delaney and Mich Manrique .....                                                                                                         | 543 |
| <i>Pre-Data Embedding Diagnostics for Construct Validity in Multi-Domain Instruments</i>                                                                       |     |
| Youngmi Cho, Tong Wu and Hyeonjoo Oh .....                                                                                                                     | 551 |
| <i>Who Keeps Up with Whom? Assessing GenAI Text Detection Across Version Updates</i>                                                                           |     |
| Ahnaf Chowdhury Niloy, Golnoush Haddadian and Min Kyu Kim .....                                                                                                | 560 |
| <i>Towards evaluating teacher performance in a GenAI teaching simulation of a science discussion</i>                                                           |     |
| Beata Beigman Klebanov, Jamie N. Mikeska, Mengxuan Zhao, Catherine Flynn, Devon Fetrow, Shreyashi Halder, Rutuja Ubale, Tricia Maxwell and Michael Suhan ..... | 571 |
| <i>High School Students' Generative AI Over-reliance and Learning Goal Orientation</i>                                                                         |     |
| Helena Liu, Newnew Deng, Hongyi Zhai and Wenchao Ma .....                                                                                                      | 581 |
| <i>Evaluating Educators' Instructional Skills on Making Content Explicit in Instruction in a Generative AI Teaching Simulation</i>                             |     |
| Shreyashi Halder, Jamie N. Mikeska and Heather Jorgenson .....                                                                                                 | 589 |
| <i>Can Small Language Models Teach Math Through Socratic Dialogue?</i>                                                                                         |     |
| Dizhi Zhang, Mengchen Li, Zichen Wang, Ziqiao Zeng, Shruti Mehta, Talita de Paula Cypriano de Souza and Seiji Isotani .....                                    | 599 |
| <i>Using AI-Simulated Response Probabilities for Item Difficulty Estimation</i>                                                                                |     |
| Brad Bolender and Sara Vispoel .....                                                                                                                           | 606 |
| <i>Can LLMs Judge Pedagogy? Assessing Conversational AI STEM Tutoring with AI-as-a-Judge</i>                                                                   |     |
| Xintian Gao, Qian Shen and Xin Li .....                                                                                                                        | 612 |

|                                                                                                                                                                     |     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| <i>Towards Breaking the Learning System Wall Using Multimodal Tutoring Transcriptions</i>                                                                           |     |
| Danielle R Thomas, Marie Cynthia Abijuru Kamikazi, Ashish Gurung, Ishan Miglani, Shivang Gupta, Zachary Levonian, Conrad Borchers and Kenneth R Koedinger . . . . . | 622 |
| <i>Does Including Images Improve Multimodal Language Models' Accuracy on Item Discrimination Estimation</i>                                                         |     |
| Jae Jun Jong and Miryeong Koo . . . . .                                                                                                                             | 631 |
| <i>The Beta-Bernstein Bridge: Closed-Form Bayesian Measurement Beyond Monotone Item Response Functions</i>                                                          |     |
| Yumou Wei . . . . .                                                                                                                                                 | 639 |
| <i>How Reproducible Are LLM-Generated Q-Matrices? Evidence from Downstream Classification</i>                                                                       |     |
| Minkwon Kim . . . . .                                                                                                                                               | 649 |
| <i>Measuring Collaborative Reasoning with LLMs</i>                                                                                                                  |     |
| Dawei Xie, Tochukwu Eze and Marcelo Worsley . . . . .                                                                                                               | 656 |
| <i>Machine-Verifiable Item Models for Geometry Diagram Generation</i>                                                                                               |     |
| Mei Chen and Gordon Stein . . . . .                                                                                                                                 | 666 |
| <i>Beyond "Three-in-a-Row": Toward Fairer Adaptive Mastery Decisions in AI-Supported Formative Assessment</i>                                                       |     |
| Chunling Niu, Kelly D Bradley, Brian F Waltman and Biao Ma . . . . .                                                                                                | 673 |
| <i>Same Score, Same Meaning? Generative AI, Automated Scoring, and the Validity of Score Interpretations</i>                                                        |     |
| Christopher Runyon . . . . .                                                                                                                                        | 681 |