LLM Psychometrics: Applying Assessment Theories to Evaluate LLMs and LLM-Supported Learners

Giora Alexandron[1,*], Beata Beigman Klebanov[2], Jill Burstein[3], Yang Jiang[2] and Alona Strugatski[1]

[1]Weizmann Institute of Science, 234 Herzl Street, POB 26, Rehovot 7610001, Israel
[2]Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA
[3]Duolingo, 5900 Penn Avenue, Second Floor, Pittsburgh, PA 15206, USA

Abstract

This preface introduces the proceedings of the LLM Psychometrics workshop, held on April 27, 2026, in Bergen, Norway, in conjunction with the 16th International Conference on Learning Analytics & Knowledge (LAK 2026). The workshop brought together researchers and practitioners from learning analytics, educational measurement, cognitive modeling, and artificial intelligence to examine how assessment theories can inform the evaluation of Large Language Models (LLMs) and LLM-supported learners. The program combined an invited keynote, paper presentations, a poster session, and collaborative discussion. Following double-blind review, eight contributions were accepted for presentation and publication.

Keywords

Large Language Models, Assessment, Item Response Theory, Psychometrics, Responsible AI

1. Preface

Large Language Models (LLMs) are increasingly deployed to perform complex tasks that were once the exclusive domain of human learners, such as reading, writing, problem solving, programming, and data analysis. While technical benchmarks of LLM performance on such tasks dominate current evaluations [1], there has been limited research on the characteristics of the processes by which LLMs learn to perform, and ultimately perform, these tasks. Mature frameworks from educational measurement—including Item Response Theory (IRT) [2], cognitive diagnostic models such as DINA [3], and learning taxonomies—offer principled ways to study the capabilities and limitations of these systems.

Applying such theories provides insights into how LLMs differ from humans [4, 5, 6], how they align or misalign with established models of learning [7], and how assessment frameworks should evolve to remain meaningful for hybrid human+AI learners [8]. The application of Generative AI technologies to learning analytics also raises important concerns related to responsible AI, academic integrity, and misuse [9, 10].

The LLM Psychometrics workshop was established to bring together researchers from learning analytics, educational measurement, cognitive modeling, and AI to explore how theories and methods of educational assessment developed for humans can be applied to the study of LLMs and LLM-supported learners. Its goals were to advance theory-grounded evaluation, connect researchers across disciplines, and translate recurring patterns in LLM performance into responsible assessment and human–AI collaboration practices.

LLM Psychometrics Workshop, 16th International Conference on Learning Analytics & Knowledge (LAK 2026), April 27, 2026, Bergen, Norway
*Corresponding author.
giora.alexandron@weizmann.ac.il (G. Alexandron); bbeigmanklebanov@ets.org (B. B. Klebanov); jill@duolingo.com (J. Burstein); yjiang002@ets.org (Y. Jiang); alona.faktor@weizmann.ac.il (A. Strugatski)
https://www.weizmann.ac.il/ScienceTeaching/Alexandron/ (G. Alexandron); https://sites.google.com/site/beatabeigmanklebanov (B. B. Klebanov); https://sites.google.com/site/jbursteinets/ (J. Burstein); https://yangjiang.me/ (Y. Jiang); https://www.linkedin.com/in/alona-strugatski-274668256/ (A. Strugatski)
0000-0003-2676-6912 (G. Alexandron); 0000-0002-5009-3992 (B. B. Klebanov); 0000-0001-7725-7574 (J. Burstein); 0000-0002-2195-5776 (Y. Jiang); 0009-0001-3282-7067 (A. Strugatski)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-15

CEUR-WS.org/Vol-4330/preface.pdf

2. Topics of Interest

The workshop invited contributions addressing topics including:

- Application of educational measurement theories and methods to understanding AI and human–AI learning and performance.
- Systematic comparisons of human and AI learner performance.
- Systematic comparisons of human–human and human–AI team performance.
- Adaptation of conceptual frameworks for designing tasks involving AI and human–AI learners.
- Methods for detecting AI participation in complex tasks, including GenAI-enabled cheating, misuse, and implications for academic integrity.
- Methods for studying AI knowledge representation and its impact on learning and performance.
- Error analyses that provide insights into AI knowledge, skills, and learning.
- Ethical considerations and limitations of applying human-based psychometric models to LLMs.

3. Workshop Format and Audience

The half-day workshop combined an invited keynote, two presentation sessions, a poster session, and collaborative discussion. It fostered cross-disciplinary dialogue among researchers and practitioners working in psychometrics, learning analytics, cognitive modeling, and AI from both theoretical and applied perspectives.

Details about the call for papers, program, and participants were published on the workshop website at https://llm-psychometrics-lak26-plum.vercel.app/. Accepted papers are published in the open workshop proceedings via CEUR-WS.

4. Workshop Program

09:00–09:15	Welcome and overview of workshop goals; introduction of organizers and agenda; brief participant introductions and ice-breaker.
09:15–10:00	Invited keynote by Alina von Davier, Chief of Assessment at Duolingo, followed by Q&A.
10:00–11:00	Paper presentations, Session 1.
11:00–11:15	Coffee break.
11:15–12:00	Paper presentations, Session 2.
12:00–12:30	Poster session and discussion into lunch.

5. Call for Papers and Review Process

The submission deadline was December 18, 2025. Following double-blind peer review, acceptance decisions were communicated to authors on January 15, 2026. Eight contributions were accepted for presentation and publication.

6. Accepted Papers

The accepted contributions are listed below in the presentation order published on the workshop website:

- *Assessing the Ability of Large Language Models to Give Learning Recommendations with Knowledge Space Theory* — Peter Steiner and Jan Hochweber.

- *From Proprietary to Open-Source: A Comparative Evaluation of LLMs for Automatic Feedback Generation* — Elisabetta Mazzullo and Okan Bulut.
- *Exploring Evaluation Methods for Generative AI-Powered Chatbot Output* — Magdalen Beiting-Parrish and Jodi Casabianca.
- *Toward Measurement Equivalence: LLM-Powered Translation of Critical Thinking Assessments* — Euigyum Kim, Hyo Jeong Shin, Alina A. von Davier, and Salah Khalil.
- *Assessment Design in the AI Era: Applying Psychometric Theory to Identify Items on Which Humans and Chatbots Diverge* — Licol Zeinfeld, Alona Strugatski, Ziva Bar-Dov, Ron Blonder, Shelley Rap, and Giora Alexandron.
- *Correcting Human Labels for Rater Effects Using Item Response Theory Models* — Jodi Casabianca and Magdalen Beiting-Parrish.
- *Psychometric Analysis of LLM Differential Sensitivity to Word Predictability using Explanatory MFRM* — Wesley Morris, Langdon Holmes, and Scott Crossley.
- *From Leaderboards to Decisions: A Psychometric Perspective on Criterion-Referenced Evaluation of Large Language Models* — Chengyuan Yao, Zhen Xu, Jiayu Zheng, and Renzhe Yu.

7. Keynote

The invited keynote was delivered by Alina von Davier, Chief of Assessment at Duolingo. The keynote and subsequent Q&A situated the workshop's central questions within contemporary assessment practice and the rapidly evolving role of LLMs in education.

8. Organizing Committee

- Giora Alexandron, Associate Professor, Weizmann Institute of Science.
- Beata Beigman Klebanov, Principal Research Scientist, Educational Testing Service.
- Jill Burstein, Principal Assessment Scientist, Duolingo.
- Yang Jiang, Research Scientist, Educational Testing Service.
- Alona Strugatski, Postdoctoral Researcher, Weizmann Institute of Science.
- Licol Zeinfeld, MSc Student and Graduate Researcher, Weizmann Institute of Science; webmaster and proceedings chair.

9. Program Committee

We would like to express our sincere gratitude to the members of the program committee for contributing their time and expertise to the review process:

- Yael Feldman-Maggor (Ben-Gurion University of the Negev).
- Elad Yacobson (Technion–Israel Institute of Technology).
- Tanya Nazaretsky (Swiss Federal Technology Institute of Lausanne).
- Meirav Attali (Taldor Assessment).
- Michael Ilagan (University of Alberta).

Declaration on Generative AI

During the preparation of this preface, the authors used ChatGPT for grammar and spell checking.

References

[1] B. Borges, N. Foroutan, D. Bayazit, A. Sotnikova, S. Montariol, T. Nazaretsky, M. Banaei, A. Sakhaeirad, P. Servant, S. P. Neshaei, J. Frej, A. Romanou, G. Weiss, S. Mamooler, Z. Chen, S. Fan, S. Gao, M. Ismayilzada, D. Paul, P. Schwaller, S. Friedli, P. Jermann, T. Käser, A. Bosselut, EPFL Grader Consortium, EPFL Data Consortium, Could ChatGPT get an engineering degree? evaluating higher education vulnerability to AI assistants, Proceedings of the National Academy of Sciences 121 (2024) e2414955121. doi:10.1073/pnas.2414955121.

[2] R. J. de Ayala, The Theory and Practice of Item Response Theory, Guilford Publications, 2013.

[3] J. de la Torre, The generalized DINA model framework, Psychometrika 76 (2011) 179–199. doi:10.1007/s11336-011-9207-7.

[4] A. Strugatski, G. Alexandron, Applying IRT to distinguish between human and generative AI responses to multiple-choice assessments, in: Proceedings of the 15th International Learning Analytics and Knowledge Conference, Association for Computing Machinery, 2025, pp. 817–823. doi:10.1145/3706468.3706490.

[5] L. Zeinfeld, A. Strugatski, Z. Bar-Dov, R. Blonder, S. Rap, G. Alexandron, Identifying items on which humans and chatbots diverge using differential item functioning, in: A. Botelho, M. M. T. Rodrigo, A. Singla, H. Ogata, H. So, Y. H. Cho (Eds.), Proceedings of the 19th International Conference on Educational Data Mining, International Educational Data Mining Society, Seoul, Republic of Korea, 2026, pp. 479–486. URL: https://educationaldatamining.org/edm2026/proceedings/2026.EDM.short-papers.88/index.html. doi:10.5281/zenodo.21039815.

[6] A. Strugatski, L. Zeinfeld, J. Cooper, S. Rap, G. Schwarts, G. Alexandron, Interpretable humans, alien LLMs: Expert analysis of latent structures in assessment responses, in: J. Wilson, C. Ormerod, M. Beiting-Parrish (Eds.), Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers, National Council on Measurement in Education (NCME), Wyndham Grand Pittsburgh Downtown, Pittsburgh, Pennsylvania, United States, 2026, pp. 303–313. URL: https://aclanthology.org/2026.aimecon-main.34/.

[7] Y. Jiang, J. Hao, M. Fauss, C. Li, Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native english speakers?, Computers & Education 217 (2024) 105070. doi:10.1016/j.compedu.2024.105070.

[8] B. G. Schelble, C. Flathmann, N. J. McNeese, G. Freeman, R. Mallick, Let's think together! assessing shared mental models, performance, and trust in human-agent teams, Proceedings of the ACM on Human-Computer Interaction 6 (2022) 1–29. doi:10.1145/3492832.

[9] M. Khalil, E. Er, Will ChatGPT get you caught? rethinking of plagiarism detection, in: Learning and Collaboration Technologies, Springer Nature Switzerland, 2023, pp. 475–487. doi:10.1007/978-3-031-34411-4_32.

[10] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martínez-Maldonado, G. Chen, X. Li, Y. Jin, D. Gašević, Practical and ethical challenges of large language models in education: A systematic scoping review, British Journal of Educational Technology 55 (2024) 90–112. doi:10.1111/bjet.13370.