Impactful and Responsible AI Systems for Education Workshop at the Festival of Learning 2026

Debshila Basu Mallick DEBSHILA@RICE.EDU
SafeInsights–OpenStax, Rice University
Simon Woodhead SIMON.WOODHEAD@EEDI.COM
Eedi Labs, United Kingdom
Zichao Wang JACKWA@ADOBE.COM
Adobe Research
Muktha Ananda MUKTHANANDA@GOOGLE.COM
Google
Jill Burstein JILL@DUOLINGO.COM
Duolingo
April Murphy AMURPHY@CARNEGIELEARNING.COM
Carnegie Learning

IRAISE 2026: The Workshop on Impactful and Responsible AI Systems for Education builds on two prior workshops held at the AAAI Conference on Artificial Intelligence. The first was AI for Education: Bridging Innovation and Responsibility (PMLR Volume 257). The second was Innovation and Responsibility in AI-Supported Education (PMLR Volume 273). This third volume collects the peer-reviewed contributions presented at IRAISE 2026, the workshop on Impactful and Responsible AI Systems for Education (IRAISE). The workshop was held on June 28, 2026 at the COEX Convention and Exhibition Center in Seoul, South Korea. It took place as part of the Festival of Learning 2026, which co-located three events: the International Conference on Artificial Intelligence in Education (AIED), the International Conference on Educational Data Mining (EDM), and the ACM Conference on Learning at Scale (L@S)

Across these editions, the community has returned to a persistent tension: the speed of AI innovation continues to outpace the slower, evidence-based rhythm of educational change. IRAISE 2026 was organized to confront that tension directly, and, as this preface aims to show, the community responded with evidence rather than positions.

1. Three pillars

The workshop was organized around three complementary pillars that we believe are necessary to build AI systems that are both impactful and responsible in education. The first, *from static to continuous*, calls on the field to move beyond one-time evaluations toward continuous improvement cycles that can keep pace with rapidly changing AI systems while respecting the deliberate timelines of educational research. The second, *from tech-first to learning theory-driven*, insists that AI tools be grounded in learning science, cognitive science, and psychometric theory, so that technical sophistication is matched by pedagogical soundness rather than assumed to imply it. The third, *from labs to classrooms*, treats the

translation of research into deployable products as a first-class problem that requires multi-stakeholder co-design with teachers, students, families, policymakers, and communities. The pillars structured the call for papers, the invited program, and the roundtables, and the next two sections record, respectively, how the day was built and what the assembled work demonstrated.

2. A program built for argument

Two format choices distinguished this edition. The first was the paired invited talks. Rather than individual speakers presenting their own work in isolation, each invited session brought together collaborators who hold complementary vantage points on a single project. These pairings took several forms: the organization building a system alongside the organization evaluating it, the researchers probing model capabilities alongside the platform supplying authentic student data, or the designers of an intervention alongside the learning scientists studying its use in context. The keynote followed the same logic, pairing a foundation model provider with an edtech evaluation lab, so the audience heard the build decisions and the evidence in the same session, with the tensions between them visible rather than smoothed away in a single narrator's account.

Muktha Ananda and Simon Woodhead opened the day. The keynote, *From exploratory signal to scale-up: evaluating a misconception-grounded AI tutor*, was delivered by Kevin McKee (Google DeepMind) and co-authored with Bibi Groot (Eedi Labs). Bibi was unfortunately unable to attend the workshop in person due to exigent circumstances. Four paired invited talks followed: *Responsible AI in Educational Assessment*, by Kevin Yancey (Duolingo) and Diego Zapata-Rivera (ETS), with Ikkyu Choi (ETS); *Can Large Language Models Understand How Students Learn?*, by Shashank Sonkar (University of Central Florida), Eamon Worden (Worcester Polytechnic Institute), and Neil and Cristina Heffernan (ASSISTments); *Evaluating MATHstream as a supplement to a blended, core middle school math curriculum*, by Stephen Fancsali (Carnegie Learning) and Ana Ribeiro (SCALE Lab, Stanford University); and *Learning Isn't Neutral: Designing Context-Aware AI for Teaching and Learning*, by Temple Lovelace (Assessment for Good and Oluko Learning) and YJ Kim (University of Adelaide). Ten contributed poster spotlights of 90 seconds each previewed the contributed program, and three poster sessions, held during the morning break, the lunch period, and the afternoon break, gave authors and attendees extended time for networking.

The second format choice was the afternoon's small-group roundtables, designed to surface tensions rather than manufacture consensus. Participants self-selected into one of four topics, *agentic AI safety, multimodal assessment, co-design methods, and bridging research-practice gaps*, and each table received not a theme but a motion, a deliberately provocative claim to argue: that agents acting on a learner's behalf have no place in classrooms until they can be continuously monitored; that multimodal assessment is surveillance dressed as pedagogy; that authentic co-design is incompatible with how educational technology is actually built and funded; and that the research-practice gap is not a translation problem but a matter of misaligned incentives. Tables opened with silent written reactions to capture the quiet voices before the loudest ones set the frame. Facilitators were briefed to surface disagreements, to have participants articulate the strongest version of the other side before

critiquing it, and to protect airtime for students and practitioners. A dedicated scribe at each table captured verbatim lines along with a structured record of three things. These were the sharpest disagreement with both sides named, the open problem the field must solve, and one path toward reconciliation. That last field was defined deliberately: not a solution the table agreed on, since the disagreement was expected to survive the session, but the move that would advance it, whether evidence that would change minds, a test both sides would accept as fair, or a safeguard that lets both positions coexist while the question stays open.

On the day of the workshop, these small-group discussions ran for nearly an hour, long enough to move past opening positions into sustained argument, before the closing session brought the tables together. There, each table's elected voice presented the key highlights of its roundtable to the whole room, and, with Debshila Basu Mallick moderating, invited everyone present to chime in and to argue for or against the points raised. Rather than ending in summary, the session concluded by naming the key actions that participants in the room, and others across the community, should take before the workshop reconvenes. The question holding the session together was what the field would need to collectively work together on by IRAISE 2027 for each disagreement to have moved: not solved, but moved. Basu Mallick closed the workshop with the next steps, and the four sets of actions that emerged from the conversations.

3. What the work showed

Read together, the keynote, the invited talks, and the contributed papers do more than sample the field; they show each of the three pillars already in motion, and they show where each one strains.

On the shift from static to continuous, the keynote modeled what a moving evidence base looks like. McKee described a collaborative project between Google DeepMind and Eedi Labs featuring a two-stage pipeline. First, an exploratory three-arm trial demonstrated that their supervised AI math tutor matched or exceeded qualified human tutors in immediate learning and applying knowledge to new topics. This feeds into a confirmatory four-arm randomized trial of roughly 1,200 secondary students across 10 English schools, designed to isolate what rich personalization adds on top of a strong pedagogy prompt. Fancsali and Ribeiro traced the same staged logic at Carnegie Learning, moving from matched correlational evidence linking MATHstream usage to end-of-year test performance toward a large-scale randomized controlled trial with Stanford's SCALE Lab. The contributed papers, meanwhile, supplied the instruments on which continuous evaluation will run. Bulut and Walsh (2026) built a seven-agent quality assurance system that evaluates LLM-generated feedback against six criteria and autonomously revises it until it passes, approving 99 percent of feedback within three iterations and escalating 1 percent to human review, where a single-agent baseline escalated 68 percent. Croteau and Heffernan (2026) argued that when a stronger model refreshes a benchmark, the shrinking set of still-failed items should be audited rather than tallied. Revisiting a 376-item image-required mathematics benchmark under a newer model, they triaged the residual failures and found that most were not model weaknesses at all but inaccessible images, under-specified problems, or benchmark defects, with only a minority worth probing as genuine visual-reasoning failures. Moreau-Pernet

(2026) released an open framework for evaluating whether simulated learners behave like real ones, scoring simulations against distributions drawn from real tutoring conversations, and Louie et al. (2026) showed that evaluation itself can move at product speed, using a micro-randomized trial that randomized the design of post-practice review after every AI-simulated therapy session and identified AI feedback paired with utterance-level reflection as the strongest condition. Li et al. (2026), reviewing 107 empirical studies of automated scoring, found prompt-based LLMs approaching parity with fine-tuned transformers while the evaluation practices around them remain inconsistent.

On the shift from tech-first to learning theory-driven, the strongest pattern in the volume is the field routing new AI through old measurement theory, and the theory not flattering the AI. Worden, Sonkar, and the Heffernans introduced FoundationalASSIST, a dataset of 1.7 million interactions from 5,000 students built precisely so that language models can be tested against learning constructs rather than correctness labels, and reported that four frontier models barely exceed trivial baselines on knowledge tracing and fall below random chance on item discrimination. Sharma (2026) subjected MRBench V2, itself an evaluation instrument for LLM tutors, to a formal psychometric pipeline and found that two of its eight dimensions degrade the scale and that measurement invariance fails between large and small models, meaning the instrument functions differently across the very comparison it is used to make. Das (2026), drawing on a decade of classifying more than 3,000 examination items with Bloom's taxonomy, named the risk of digital rote, AI reproducing pedagogically shallow learning at scale, and proposed a cognitive audit, specification, and validation cycle analogous to a fairness audit. Constructive versions of the same instinct appeared throughout: Ilídio et al. (2026) extended Duolingo's half-life regression with learner and item features that recover interpretable ability and difficulty parameters for spaced repetition; Mitton et al. (2026) showed that uncertainty-aware knowledge tracing models can learn when to defer to a teacher, and asked openly whether model uncertainty is rediscovering classical item response theory; Scaria et al. (2026a) argued in a position paper for assessment-centered, bounded pedagogical agents whose progression decisions rest on structured evidence of learner performance rather than open-ended chat, with the same group's HeteroForge system (Scaria et al., 2026b) applying multi-agent debate to curriculum-aligned exercise generation; and Jin and Chen (2026) kept rubric design under instructor control by simulating the distributional consequences of rubric choices before any live grading.

On the shift from labs to classrooms, the volume's classroom studies earned their evidence in authentic conditions and reported it candidly. Heickal and Lan (2026) randomized feedback conditions within a real introductory Python course and found that natural language hints roughly doubled the odds of eventual success on a problem, while immediate per-attempt gains were near zero across all conditions, a distinction only a longitudinal classroom deployment could surface, and released the resulting dataset. Brender et al. (2026) ran a semester-long field study in a 172-student robotics course of an AI arguing agent that deliberately challenges students with common misconceptions, and reported an honest null: no significant differences in error correction between peer-peer and peer-AI argumentation, with students rating the AI partner below a human peer. Lovelace and Kim made context the design target itself, showing through the New Insights and CRP-JeepyTA cases how systems can make learners' resources and teachers' contexts salient rather than flattening them, and Lee et al. (2026a) documented what co-design actually looks like from inside, a

year-long, nonlinear negotiation of authority among the stakeholders of a multimodal learning analytics platform for secondary STEM classrooms. Notably, translation arrived with governance attached. Jeon et al. (2026) surveyed educators across nine districts and found support concentrated on teacher-mediated uses of AI, with the strongest concerns about erosion of student reasoning and the strongest demands for formal oversight frameworks; Shi and DiFranzo (2026) proposed bounded confidentiality as a middle path between the two bad defaults of total surveillance and total secrecy when tutoring conversations become disclosure sites; Xiang et al. (2026) built classroom engagement analytics that decouple identity from signal by design, reporting only region-level patterns to a teacher dashboard; and Lee et al. (2026b) analyzed over 100 responses to a public request for information on the datasets, benchmarks, and models the field needs as shared infrastructure, echoed at the organizational level by Yancey's account of how Duolingo's cross-disciplinary team structure produces its responsible AI standards and by Zapata-Rivera and Choi's survey of ETS practice across content generation, conversation-based assessment, scoring, security, and score reporting. Human review remained load-bearing rather than decorative across the volume, from Badea and Dumitran (2026), whose two-validator annotation platform routes every AI-generated geometry description through two independent human passes, to Moroianu et al. (2026), whose editorial-generation workflow treats every LLM draft as material for human refinement, to Wei et al. (2026), who found that adding context to AI-assisted discovery in educational data archives can help and harm, shifting the model's reasoning mode rather than simply improving it.

Three patterns cut across the pillars and are worth naming because they mark a maturing field. First, disconfirming evidence was presented as contribution rather than failure: the null result in a real course, the frontier models below chance on item discrimination, the near-zero immediate gains behind a real longitudinal effect. Second, knowing when not to act emerged as a design principle, in deferral to teachers under uncertainty, in bounded agents and bounded confidentiality, in escalation to human review, and in audit queues; the responsible systems in this volume are distinguished less by what they automate than by what they decline to. Third, the human in the loop appeared as architecture rather than afterthought, with validators, instructors, and teachers holding decision rights by design. These patterns did not settle the field's arguments; they sharpened them, and the roundtables took those arguments up directly.

4. Before we meet again in 2027

IRAISE has always been intended as a community rather than a single event, and the round table's closing question, what must be true by IRAISE 2027 for each disagreement to have moved, gives this preface its natural ending. We close by recording where the arguments landed and the specific work that would move them.

On agentic AI safety, the table split over whether autonomous agents belong in classrooms at all before the safeguards exist. One side held that agentic AI should be paused or kept out until continuous monitoring, clear guardrails, privacy protections, and enforceable accountability are in place, since without them students may lose agency, harmful guidance may go undetected, and vendors may externalize the risks while pursuing profit. The other side held that agentic AI is already becoming inevitable, that waiting for perfect monitoring

could cause students and teachers to lag as the technology advances, and that imperfect, bounded deployment under teacher oversight may be safer and more realistic than prohibition, particularly when students feel more comfortable asking an AI than asking a teacher. The open problem underneath is twofold: what level of oversight is sufficient to let an agent act for or advise a learner without removing the learner's agency or placing an impossible monitoring burden on teachers, and what accountability model should govern harm when it occurs, whether responsibility lies with the vendor that built and sold the system, the school that procured it, the teacher overseeing its use, the government that set or failed to set standards, or some combination. The path forward the table proposed is a staged, bounded deployment test that both sides could accept: before an agent interacts directly with students it should be evaluated against human tutors in controlled trials, tested for harmful guidance, bias, privacy risk, and loss of learner agency, and restricted to clearly defined tasks. The table named who acts at each layer: vendors supply evidence about system limits, monitoring, data use, and failure modes; schools and teachers define where human review is mandatory; governments or independent regulators set universal minimum standards, liability rules, and audit requirements; and students retain the right to know when an agent is acting for them, to review its actions, and to refuse its use. Such a test would not settle whether agentic AI belongs in classrooms, but before 2027 it would produce the evidence and safeguards that allow cautious experimentation without treating students as an uncontrolled test population.

Debates around multimodal assessment split along two main lines: first, whether trustworthy assessment is possible without multimodal signals; and second, whether multimodal systems are effective at all—given that determined cheaters adapt while surveillance induces stress and false positives among honest students. Beneath this divide lies a critical ethical question: can piloting unvalidated systems on real students ever be justified? To chart a path forward, the group rejected a one-size-fits-all surveillance model in favor of three coexisting moves: redesigning assessment formats to outpace AI assistance, strictly anonymizing and aggregating collected data, and reserving multimodal systems for contexts where students directly benefit and are able to consent and inspect their own data (e.g., low-stakes, self-motivated learning). Moving forward, the community must articulate which explicit signals of learning and engagement (e.g., assessment completion and performance, course grades, self-reports) are defensible for data collection, clarify where implicit ones are not (e.g., body language, eye movements, collaborative group dynamics), and establish solid validity evidence prior to deployment.

On co-design methods, the table split over whether authentic co-design is feasible from the start of a product's life or only under research conditions. One side held that co-design is possible from day one given the right resources and is core to the design process, as it is in many research settings. The other side held that it is difficult to reconcile product intuition, innovation, and fast iteration with genuine co-design alongside teachers, particularly in a startup environment. The open problem is the absence of incentives. It is hard to demonstrate the value of co-design in the finished product. Although research settings have the resources and long timelines to co-design solutions, those tools rarely become available outside the research context. Even meeting teachers can demand significant effort that is not feasible to sustain regularly. The identified path forward has two parts. The first is evidence of the value of co-design for companies and products, so that value can function as

the missing incentive. The second is a sharper definition of co-design itself, one that does not place the teacher and the designer at the same level. Teachers know what is good for them and can surface pain points but do not necessarily know how to design the product, while designers should arrive without preconceived solutions, listen, and translate what teachers say into actionable product requirements that address the underlying issue rather than taking the stated request at face value. Before 2027, teams can begin publishing that evidence and adopting this translational definition of the designer's role, so that co-design becomes a documented practice with demonstrated product value rather than a label.

On bridging research and practice, two splits ran through the table. The first set standardization against individualization. One camp held that scale requires common standards and common evidence, and the other that the value of these systems lay precisely in classroom-level personalization that resists standardization. The second, left unresolved, is who the customer even is: a service sold to teachers or a product sold to administrators, who buy for different reasons. The open problem is how to scale an intervention whose effectiveness depends on adapting to individual classrooms, and how validated research finds uptake when there is little demand-side pull for it. The table's identified path forward is a shared effectiveness test that both camps accept, with the metric agreed in advance and with explicit treatment of what counts as evidence of adaptation versus evidence washed out by adaptation, designed around the reality that technology development cycles run faster than randomized trials, so evidence will arrive later than the products it is meant to evaluate. The table also named who acts, specifically, public bodies to supply the grounding that is currently absent, and organizations to close the demand gap so that buyers know the research exists.

Readers who wish to revisit the program or follow the next edition can visit https://iraise2026.org, and those who would like to take up any of these threads, contribute evidence that would change minds, propose a test both sides would accept, or help shape the 2027 program are warmly invited to reach the organizers at iraise-26-fol@googlegroups.com. We hope this volume is useful to researchers and practitioners alike, and we look forward to seeing which of these disagreements have moved, not solved but moved, when the community reconvenes in 2027.

5. Submissions and review

All submissions were managed through OpenReview and underwent double-blind peer review, with full papers reporting original research and short papers covering demos, work in progress, position papers, and practitioner reports, all in the PMLR style. The workshop received **32** submissions and accepted **23** papers (**15** full and **8** short), reviewed by **N** program committee members. One accepted short paper was not submitted for publication, so this volume collects the remaining **22** papers (**15** full and **7** short). Accepted papers were invited to submit extended versions addressing reviewer remarks for publication in this volume.

Table 1: Summary of the IRAISE 2026 papers in this volume

Format	Oral	Poster	Total
Full	8	7	**15**
Short	2	5	**7**
Total	**10**	**12**	**22**

Acknowledgments

A workshop of this scope is the product of many hands. We thank the members of the IRAISE 2026 program committee, whose careful and timely reviews under a compressed timeline made the technical program possible. We are grateful to our keynote speaker, Kevin McKee, and to our invited speakers, Kevin Yancey, Diego Zapata-Rivera, Shashank Sonkar, Eamon Worden, Neil Heffernan, Cristina Heffernan, Stephen Fancsali, Ana Ribeiro, Temple Lovelace, and YJ Kim, for talks that set a high standard for evidence and candor. We thank the roundtable facilitators, Simon Woodhead, Muktha Ananda, Jeremy Roschelle, YJ Kim, and Seungho Jeon, for cultivating safe arguments and dialog; the scribes, Nicy Scaria, Sabin-Codruţ Badea, Helen Jin, Baptiste Moreau-Pernet, and Gabriel Diaz, whose live capture of the tables' sharpest exchanges made the capstone and the closing agenda possible; the elected table voices who carried their groups' arguments onto the stage; and the authors and participants whose contributions and questions were the substance of the day and the preface. We recognize the recipient of the IRAISE travel scholarship sponsored by Eedi, Nicy Scaria, who was able to join us and enrich our conversations and community. Finally, we thank the Festival of Learning 2026 organizers and the local team in Seoul, together with the communities behind AIED, EDM, and L@S, for hosting the workshop, and we thank the series editors of the Proceedings of Machine Learning Research for supporting the publication of this volume.

References

Sabin-Codruţ Badea and Adrian Marius Dumitran. A two-validator web interface for structured geometry figure annotation. pages 107–111.

Jérôme Brender, Aitor Perez, Patrick Jermann, Engin Bumbacher, Francesco Mondada, and Chenyang Wang. More correcting or confounding? When an AI arguing agent informed by common student misconceptions joins case-based peer argumentation in a real course. pages 80–89.

Okan Bulut and Cole Walsh. A multi-agent system for feedback quality assurance: Evaluate, revise, and validate at scale. pages 1–11.

Ethan Croteau and Neil Heffernan. Beyond aggregate accuracy: Continuous evaluation of residual failures in image-required educational mathematics. pages 170–179.

Syaamantak Das. Cognitive grounding before AI scaling: Why AI-in-education systems need learning theory first. pages 129–139.

Xin Wei, Morgan Lee, Jie Min, and Jeremy Roschelle. More sources, better AI? When context helps and harms AI-assisted discovery in educational data archives. pages 112–122.

Zhiyuan Xiang, Qiao Lin, Xueyan Jia, and Shengkai Zheng. Privacy-preserving region-level classroom engagement analytics for teacher reflection. pages 160–169.