

Principled Approaches to Building AI Partners for Assessing and Supporting Small-Group Collaboration in Classrooms

Peter W. Foltz, Chelsea Chandler, Mon-Lin Monica Ko

Institute of Cognitive Science, University of Colorado Boulder

foltzp@colorado.edu

Abstract

Collaboration is a complex, multifaceted construct, blending cognitive, social, and emotional components; it is context-dependent, dynamic, multimodal and multiparty, properties that resist simple, one-dimensional measurement. We present a principled framework for addressing this complexity that links what is measured, where it is measured, and how evidence is calculated and warranted. This paper (1) synthesizes a taxonomy of key design dimensions for building AI partners that support small-group collaboration; (2) demonstrates how it operationalizes these dimensions in practice. to show how design choices inform how measurement science can be embedded in AI-based collaboration partners.

1 Introduction

As classrooms and workplaces increasingly organize around joint problem-solving rather than individual work, the capacity to collaborate effectively has become one of the most consequential skills a learner can develop. Practicing collaborative problem solving (CPS) helps develop critical thinking, communication, and leadership skills while boosting student engagement (Graesser et al., 2018; Springer et al., 1999; Xu et al., 2023), and employers increasingly rank collaboration alongside problem solving and adaptability as core to career resilience (World Economic Forum, 2025). Unlike many constructs targeted by AI-based assessment, collaboration is a complex, multifaceted construct that blends cognitive, social, and emotional components; it is highly context-dependent; it unfolds dynamically over time rather than as a single snapshot; and it is inherently multimodal and multiparty (Blikstein & Worsley, 2016). These properties make collaboration resistant to one-dimensional measurement and require AI-based approaches

capable of modeling nuanced, evolving interpersonal interaction.

Much of the recent focus around AI in education has centered on personalized, one-on-one tutoring, exemplified by large language model (LLM)-based tutors that adapt to an individual learner's needs. Yet learning is not only personal; it is also a social activity situated within a cultural system (Greeno, 1998), and collaborative, open-ended learning environments are critical for fostering deeper engagement and knowledge building (Jeong et al., 2019). Building AI partners that meaningfully support, rather than merely observe, group collaboration therefore requires design principles distinct from those governing single-user tutoring systems.

Evidence-centered design (ECD; Mislevy et al., 2006) provides a principled framework for addressing this complexity by explicitly linking what is measured, where it is measured, and how evidence is calculated and warranted. Applied to collaboration, it clarifies the link between observed discourse or behavior and claims about underlying skills, helping prevent the common failure mode of AI systems optimizing for what is easy to capture (e.g., word counts, click rates) rather than what is construct-relevant (e.g., negotiation, knowledge transfer, perspective-taking). It also supports transparency: by mapping behavioral markers explicitly back to specific evidentiary claims, it reduces the “black-boxiness” of AI-based assessment and supports stakeholder trust among learners, educators, and policymakers.

This paper has three aims. First, we synthesize a taxonomy of key dimensions that guide the principled design of AI partners for collaboration in classrooms (Foltz, 2026). Second, we demonstrate how ECD operationalizes these dimensions in practice. Third, we illustrate these principles by synthesizing work from two real-world implementations developed at the NSF National AI Institute for Student-AI Teaming (iSAT) (D'Mello et al., 2024): the Community

Builder (CoBi; Breideband et al., 2023), which supports relationship-building and socially uplifting discourse during open-ended small-group science tasks, and the Jigsaw Interactive Agent (JIA; Ko et al., 2026), which supports productive uncertainty during structured jigsaw collaboration. They illustrate a range of design choices, including AI models, aggregated versus individualized feedback, and varying timing and intervention strategies, that inform how measurement science can be embedded into the next generation of AI-based collaboration partners.

2 Frameworks for Characterizing Collaboration Skills

Assessing collaboration requires identifying the skills to be targeted. Several frameworks have been developed for this purpose, each emphasizing different aspects of cognitive and social functioning. The ATC21S project (Griffin et al., 2012) characterized CPS along participation, perspective-taking, social regulation, and negotiation, focusing on authentic human-to-human interaction but relying on skill descriptors that are difficult to automate. OECD's PISA 2015 Framework (OECD, 2017) defined CPS along a cognitive dimension (task work: exploring, representing, planning, monitoring) crossed with a collaborative dimension (teamwork: establishing shared understanding, taking action, organizing the group), that was scored automatically but within highly scripted, computer-agent interactions. Sun et al. (2020) extended this approach with a focus on joint analysis, negotiation, and knowledge construction, defined via specific, measurable discourse indicators (what is said and done) that support natural language analysis of open, unconstrained conversation.

Despite their differences, these frameworks converge on a common underlying structure: successful collaboration requires both task work (problem-solving strategies, planning, monitoring) and teamwork (communication, coordination, relationship management). Frameworks that link skills to computationally tractable, discourse-level indicators are considerably more amenable to automated assessment than those defined at a broad, human-scored level. What is needed beyond a skills framework, however, is a principled approach for mapping those skills to evidence,

3 Evidence-Centered Design for AI-Based Collaborative Assessment

Evidence-centered design (ECD; Mislevy et al., 2006) provides a comprehensive framework for designing and deploying multimodal, multiparty AI-based assessments by explicitly outlining what is measured (the student model), where it is measured (the task model), and how evidence is calculated (the evidence model). ECD begins with a student model, the claims one wants to make about a learner's or group's knowledge, skills, or abilities, and works backward to define the evidence needed and the tasks that can elicit it. For AI-based assessment, this backward-design logic prevents the common problem of measuring what is easy to capture rather than what is construct-relevant, keeping the analysis focused on construct-relevant features rather than spurious correlations (Foltz, 2026).

Because collaborative data are inherently multimodal, spanning speech, text, gesture, and action logs, ECD requires specifying ahead of time which modalities are relevant in a given context, which features to extract, and how those features support inferences about the target constructs. ECD also enforces explicit reasoning about why a given behavior or pattern of interaction provides evidence for a construct, which removes some of the "black-boxiness" of an AI assessment system: as an AI detects individual markers of behavior, these markers can be explicitly mapped back to specific pieces of evidence, providing a transparent accounting of the analysis process that is essential for stakeholder trust. Prior ECD-grounded implementations illustrate this approach at different scales. For example, the in-task assessment framework (I-TAF; Kerr et al., 2016) applies a CPS ontology to extract behavioral evidence from log data such as mouse clicks and chat messages, while the EPCAL platform (Hao et al., 2017) was built to support control and manipulation of group variables for studying collaboration at scale. As we describe in this paper, we extend this approach with partners to support real-time, open-ended, spoken small-group discourse in authentic K–12 classrooms.

4 A Taxonomy of Design Dimensions for AI-Based Collaborative Support

Identifying collaboration skills and mapping them to evidence is necessary but not sufficient for

building AI technology to assess and support collaboration: these design choices cannot be evaluated in isolation from the practical contexts in which they are applied. Table 1 (adapted from Foltz, 2026), enumerates the key dimensions to consider when designing and deploying AI-based collaborative assessment and support tools, organized into five domains, context and setup, modeling and evidence, timing, feedback and use, and scope and scale, along with each dimension's primary alignment to the ECD student, task, and evidence models, or to practical implementation factors. We summarize each domain below.

4.1 Context and Setup

The context and setup domain addresses the foundational conditions under which an assessment is designed and conducted. The purpose of assessment, whether formative, summative, or diagnostic, carries implications for how the system is structured and what feedback is provided; a formative purpose, in particular, requires mapping identified skills to specific, pedagogically effective feedback. The interaction environment (simulated, live, or hybrid) shapes both the realism and the controllability of the collaboration: live classroom interactions capture authentic dynamics but introduce noise, interruptions, and off-task behavior, whereas simulated settings offer tighter control at the cost of naturalistic behavior. Data-collection choices (spoken audio, video, log files, final products) depend on feasibility and on the richness of evidence required, while also introducing privacy considerations that are heightened when working with minors. Task structure and control range from tightly scripted tasks to open-ended conversation, influencing the range of competencies that can be observed. Finally, teammate embodiment, whether the AI acts as a virtual teammate, provides feedback only, or supports a human-in-the-loop, affects student engagement, trust, and interaction patterns.

4.2 Modeling and Evidence

The modeling and evidence domain concerns what is measured and how evidence is captured, processed, and interpreted. Targeted competencies may include CPS skills, socio-emotional skills, pedagogical moves, or domain-specific knowledge, and can be assessed at varying levels of granularity, including individual, dyad, or whole

team. Evidence sources range from verbal and non-verbal discourse markers to group work products, action logs, and observer ratings. Analytic models span a continuum from simple frequency counts to statistical and Bayesian models to fine-tuned language models and LLMs, with more sophisticated multimodal fusion approaches consistently outperforming unimodal counterparts (e.g., Cohn et al., 2024). Communication among team members remains one of the richest sources of evidence about collaborative process and cognition, but spoken discourse must first be processed by automatic speech recognition (ASR), which can exhibit word error rates of 20 to 80% in noisy, real-world classrooms with young speakers, posing a substantial challenge for downstream discourse classification (Southwell et al., 2022).

4.3 Timing and Feedback and Use

The nature and timing of AI feedback can vary widely, from identifying discrete collaborative moves to diagnosing team processes, providing domain-specific scaffolding, or generating incident reports and scores. Data collection and analysis may occur in real time, near-real time, or through post hoc analysis, with each choice trading off immediacy of feedback against scalability and model reliability; a lower-fidelity model that runs in real time may be as effective as a higher-fidelity model requiring post hoc processing. Feedback delivered concurrently with a group task can shape ongoing collaboration but risks disrupting natural interaction, whereas post hoc feedback allows deeper reflection at the cost of losing the opportunity to shape in-the-moment behavior. The target of feedback, whether individual, team, or teacher and evaluator, further shapes its tone, granularity, and actionable quality, and must be balanced against interpretability, specificity, and alignment with broader learning goals.

4.4 Scope and Scale

Scope and scale address broader implementation considerations that determine whether a system can operate effectively and ethically in its intended environment. Scalability of assessment requires planning for pilot phases, small-scale trials, and eventual large-scale deployment, each with distinct infrastructure, training, and quality-assurance needs, as well as organizational buy-in and stakeholder trust. Context sensitivity addresses whether a system is domain-general, domain-

Table 1: Key dimensions for the design of AI-based collaborative assessment and support systems. Adapted from Foltz (2026),

<i>Development Domain</i>	<i>Collaborative Assessment Dimension</i>	<i>Categories</i>	<i>Alignment to ECD Layers / Implementation Factors</i>
<i>Context and setup</i>	Purpose of assessment	Formative, summative, diagnostic	<i>Implementation</i>
	Interaction environment	Simulated, live, hybrid	<i>Task model</i>
	Data types	Spoken audio, video, log files, student responses	<i>Implementation</i>
	Task structure and control	Firmly directed by tasks (multiple choice, SJT), semi-structured, open-ended (e.g., open conversations)	<i>Task model</i>
	Teammate embodiment	Virtual teammate, human, human-in-the-loop	<i>Task model</i>
<i>Modeling and evidence</i>	Targeted competencies	CPS, social-emotional, pedagogical moves, domain knowledge	<i>Student model</i>
	Granularity of analysis	Individual, dyad, whole team	<i>Student model</i>
	Collaboration evidence sources	Group work product, verbal expressions, turn-taking, influence, inferred eye gaze, pointing, situation judgment, action logs, survey instruments, observer notes, or scores	<i>Evidence model</i>
	Analytic model(s)	Frequency counts, Statistical, Bayesian, semantic feature-based, or LLM-based	<i>Evidence model</i>
	Data integration method	Single modality, multimodal, separate, multimodal-integrated	<i>Evidence model</i>
	Data access and privacy constraints	Unrestricted, restricted, anonymized	<i>Implementation</i>
<i>Timing</i>	Timeliness of data collection and analysis	Real-time, near-real-time, post hoc analysis	<i>Implementation</i>
	Timing of AI feedback	Real-time concurrent with group task, near-real-time, post hoc	<i>Task model</i>
<i>Feedback and use</i>	Nature of AI feedback	Collaborative moves, team processes, domain scaffolding, reports, critical incidents, scoring	<i>Task model</i>
	Target of feedback	Individual, team, teachers, evaluators	<i>Task model</i>
<i>Scope and scale</i>	Scalability of assessment	Pilot, small-scale, large-scale	<i>Implementation</i>
	Context sensitivity	Domain-general, domain-specific, context-adapted, culturally appropriate	<i>Implementation</i>

specific, or context-adapted, and, critically, whether it is culturally appropriate. What constitutes productive participation, appropriate communication, or effective leadership varies substantially across individualistic and collectivistic cultures and different communication norms (Gelfand & Realo, 1999); an AI system trained primarily on Western, industrialized populations that privileges verbal contribution frequency or direct challenge as markers of collaboration quality may systematically disadvantage students from cultures in which such behaviors are read differently. Meaningful accommodation of this variation requires both technical solutions (e.g., diverse training data, domain adaptation, uncertainty quantification) and ongoing participatory collaboration with educators, students, and cultural experts throughout the design and validation process.

5 Illustrative Applications

To ground the taxonomy in Table 1, we describe two AI partners developed at iSAT. It should be noted that the partners are not presented here as

competing alternatives; they target different constructs, tasks, and classroom purposes, and were designed to complement each other within a broader ecosystem of collaboration supports. This work reflects a broader commitment to responsible innovation in AI for education, one that shifts the field's attention from personalization toward more explicit emphasis on how AI supports specific forms of learning and how it is designed, both for whom and with whom (e.g., Penuel et al., 2026). Evidence-centered design offers one concrete route to that commitment, by making the constructs, evidence, and feedback choices behind an AI partner explicit rather than incidental. From an ECD-grounded design perspective: each specifies a student model (target constructs), a task model (understanding the classroom activity and data-collection context), and an evidence model (the computational approach linking discourse to constructs). Because each system resolves the dimensions in Table 1 somewhat differently, considering them together lets us illustrate a broader range of the design choices and trade-offs from the taxonomy.

5.1 CoBi: The Community Builder

The Community Builder (CoBi; Breideband et al., 2023) is a multiparty AI partner that targets the relationship dimension of collaboration. Its student model comprises co-negotiated Community Agreements (CAs), namely being respectful, showing commitment to community, and moving and thinking forward, derived from the Sun et al. (2020) CPS framework and grounded in OpenSciEd K–12 science materials aligned with national standards. Its task model situates CoBi within open-ended, small-group science tasks: with teacher guidance, students first co-negotiate concrete examples of each CA; as they then engage in collaborative tasks, CoBi analyzes their spoken discourse in real time; after the task, teachers use CoBi’s visualizations to guide collaborative reflection; and, optionally, teachers reveal CoBi’s top-ranked “noticings,” specific positive exemplars of each CA, to support both collaborative reflection and critical AI literacy.

CoBi’s evidence model captures audio via tabletop microphones placed with each student group and analyzes the resulting speech with AI language models trained to recognize evidence of each Community Agreement, outputting a likelihood that a given utterance exemplifies that agreement. Because classroom audio is noisy and automatic speech recognition is error-prone for young speakers, CoBi’s models are trained for robustness to transcription error and restrict inference to utterances that can be classified with high confidence (Southwell et al., 2022). Consistent with its evidence and task models, results are deliberately aggregated at the classroom level rather than reported to individuals, reducing distraction and surveillance while preserving CoBi’s formative, reflective purpose. Feedback is delivered through real-time visualizations, including a growing “tree” of noticings, displayed to the whole class.

CoBi is built on a scalable, modular cloud architecture and has been deployed with hundreds of middle-school student groups across multiple school districts, requiring only a Chromebook and an affordable USB microphone per group. Preliminary classroom testing found that students responded positively to CoBi’s reflective visualizations, though when confronted with their own results, students tended to attribute shortcomings to their own collaboration behavior rather than to possible errors in CoBi’s

classification, an important consideration for how transparency and AI-literacy activities are designed (Breideband et al., 2023).

5.2 JIA: The Jigsaw Interactive Agent

The Jigsaw Interactive Agent (JIA; Ko et al., 2026) targets a different construct and a more structured task. In a jigsaw activity, students first develop individual expertise on different aspects of a topic and then work together in a small group to pool that expertise and solve a problem that requires everyone’s contribution (Aronson, 1978). Its student model comprises four discourse-based intervention types derived from research on teacher facilitation of productive uncertainty, namely support, connect, problematize, and stabilize (Watkins & Manz, 2022), reflecting a pedagogical stance that seeks to help students surface and productively work through, rather than resolve away, the non-obvious and contingent aspects of their collective sensemaking (Manz & Suárez, 2018). Its task model situates JIA within a jigsaw activity embedded in a co-designed, three-week AI-literacy curriculum unit, in which students first develop individual expertise on one of five content-moderation approaches and then return to their small group to pool that expertise and rank the approaches by effectiveness, a task that structurally requires collaboration, since no single student holds complete information.

JIA’s evidence model was initially deliberately varied to compare three dialog-policy approaches, each analyzing the immediately preceding minute of group discourse and triggering one of the four intervention types directly into the group’s chat interface: a rule-based approach that selected interventions from indicators of participation structure, task engagement, and collaboration quality; an LLM-based approach (Llama 70B) that operated directly on ASR-derived transcripts, summarizing the discussion and justifying its chosen intervention in natural language; and a hybrid approach that supplied the same LLM with both the conversational transcript and the rule-based indicators. All three approaches contended with substantial automatic speech recognition error, approximately 58% word error rate in benchmark classroom conditions.

JIA was formatively evaluated with 64 small groups of 8th-grade students across 11 classrooms in two U.S. school districts, of which 16 groups with moderate-to-high engagement were the focus

of detailed human-annotated analysis (Ko et al., 2026). The LLM- and hybrid-based approaches were significantly more likely to generate connect interventions and were judged more appropriate by human coders (82% and 87%, respectively) than the rule-based approach (63%), which defaulted to more repetitive support interventions. Students reacted more often to the LLM approach (49%) than to the rule-based (24%) or hybrid (29%) approaches and displayed a wider range of emotional responses, while connect and problematize interventions elicited more elaborated uptake than support interventions..

6 The Taxonomy's Design Space

CoBi and JIA illustrate how the dimensions in Table 1 can be resolved differently depending on a system's construct, task, and purpose. In the modeling and evidence domain, JIA's formative comparison of rule-based, LLM-based, and hybrid dialog policies shows one way to weigh interpretability and computational cost against contextual responsiveness. Its LLM- and hybrid-based approaches produced more contextually responsive, construct-aligned feedback, more often connecting ideas across group members rather than defaulting to generic encouragement, but also elicited more variable emotional reactions from students, including occasional dismissiveness alongside enthusiasm. This illustrates a trade-off inherent to the modeling and evidence domain: more flexible, generative approaches can better track construct-relevant nuance in open discourse, at some cost to predictability.

The systems also illustrate the range of choices in the feedback and use domain. CoBi's classroom-level aggregation reflects a design decision aligned with its evidence and task models: because CoBi is a reflection-support tool oriented toward the relationship dimension of collaboration, aggregating feedback across groups reduces distraction and the risk that students experience the tool as surveillance, while preserving its formative, reflective function. JIA instead delivers feedback directly to each small group in real time, functioning as an embedded teammate whose goal, supporting in-the-moment productive uncertainty, calls for timely, group-specific intervention rather than delayed, aggregated reflection. Together, the two systems show how the appropriate target of feedback follows from the construct and purpose

being assessed rather than from a single best practice.

Timing choices follow a similar logic. CoBi's visualizations update continuously but are intended primarily for use during teacher-led reflection after a task, while JIA intervenes proactively at fixed intervals during group work itself. Each choice reflects the timing and feedback domain's central trade-off between disruption and immediacy: JIA's interval-based design was intended to support students in the moment without overwhelming them, while CoBi's post-task reflection avoids interrupting collaboration but limits the chance to shape it as it happens.

Finally, the two systems illustrate different points along the scope and scale domain. CoBi has been deployed at substantial scale, across hundreds of groups and multiple school districts, on a production cloud architecture. JIA remains at the formative pilot stage, with 64 groups and detailed analysis of 16, reflecting the additional validation an actively intervening agent typically requires before wider classroom use. Despite these differences in maturity, both systems demonstrate that ECD-grounded design enables defensible, real-time analysis of noisy, multiparty classroom discourse, in part by restricting inference to high-confidence evidence and by validating models against human-annotated criterion data (Foltz, 2026).

7 Toward Validated, Ethical, and Scalable AI Collaboration Partners

The AI collaborative partners demonstrate that principled, grounded design choices, including what is measured, how evidence is extracted from imperfect multimodal data, and how and to whom feedback is delivered, materially shape both the validity and the classroom impact of AI collaboration partners. Implementing AI partners at scale also requires confronting challenges that fall outside the immediate scope of Table 1 but are inseparable from responsible implementation. Multimodal data collection, particularly of children's speech, and potentially video, gesture, or eye gaze, raises privacy and consent challenges that a purely technical framework cannot resolve on its own; informed consent must move from a binary opt-in/opt-out decision toward a continuous, measurable process that maintains transparency and accountability throughout data collection (Penuel et al., 2026). Regulatory frameworks such

as FERPA, COPPA, and PPRA impose overlapping but distinct requirements that researchers and developers must navigate to protect students' rights and dignity. Demonstrating the educational value that such tools provide, for instance through transparent, critiquable feedback in the spirit of CoBi's noticings, is likely a prerequisite for broader stakeholder buy-in.

A second, related challenge is generalizability across curricula, cultural contexts, and task structures. As emphasized in the scope and scale domain, collaboration norms are not culturally universal, and systems trained on narrow populations risk misreading, or systematically disadvantaging, students whose collaborative styles differ from those of the training population. Addressing this will require diverse training data, domain adaptation and uncertainty-aware modeling, and, above all, sustained participatory design with educators, students, and cultural experts from the earliest stages of competency modeling through validation.

Collaboration is difficult to measure precisely because it is not one thing: it is cognitive and social, momentary and cumulative, individual and collective. Evidence-centered design offers a way to hold that complexity accountable to explicit claims, tasks, and evidence rather than to whatever happens to be easiest to compute. Applying ECD to real-world AI partners, shows both the promise and the trade-offs of principled design: aggregated versus individualized feedback, different points along the spectrum from validated classifiers to more flexible but variable generative models, and reflective versus in-the-moment intervention.

As AI-based collaboration partners move from research pilots toward broader classroom use, continued attention to transparent, evidence-based, and human-centered design, anchored in frameworks, will be essential to ensuring that these systems serve, rather than undermine, the students and teachers they are meant to support. The present work contributes a generalizable taxonomy of design dimensions, grounded in measurement science, for researchers and developers building the next generation of AI-based collaborative assessment and support tools. By making explicit the choices bundled into any AI collaboration partner, choices too often left implicit, we hope to give the measurement community a shared vocabulary for evaluating, comparing, and improving these systems as they move from

classroom pilots toward broader, validated deployment.

Acknowledgments

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grants DRL 2019805 and DRL 2454151. The opinions expressed are those of the authors and do not represent the views of the NSF.

References

- Aronson, E. (1978). *The jigsaw classroom*. Sage.
- Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics: A methodological framework for research in constructivist learning. *Journal of Learning Analytics*, 3(2), 220–238.
- Breideband, T., Bush, J., Chandler, C., Chang, M., Dickler, R., Foltz, P., ... & D'Mello, S. (2023, October). The Community Builder (CoBi): Helping students to develop better small group collaborative learning skills. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 376–380). ACM. <https://doi.org/10.1145/3584931.3607498>
- Cohn, C., Davalos, E., Vatrál, C., Fonteles, J. H., Wang, H. D., Ma, M., & Biswas, G. (2024). Multimodal methods for analyzing learning and training environments: A systematic literature review. *arXiv preprint arXiv:2408.14491*.
- D'Mello S. K., Bidy Q., Breideband T., Bush J. B., Chang M. A., Cortez A., Flanigan J., Foltz P. W., Gorman J. C., Hirshfield L. M., Ko M.-L. M., Krishnaswamy N., Lieber R., Martin J., Palmer M. S., Penuel W. R., Walker L., Whitehill J. (2024). From learning optimization to learner flourishing: Reimagining AI in education at the Institute for Student-AI Teaming (iSAT). *AI Magazine*.
- Foltz, P. W. (2026). Key considerations for assessing and supporting collaboration with AI. In (X. Xiong, M. D. Shermis & J. Xiong, Eds.). *The Role of AI in Assessment* (pp. 287–311). Routledge, London. <https://doi.org/10.4324/9781003621522-12>
- Gelfand, M. J., & Realo, A. (1999). Individualism-collectivism and accountability in intergroup negotiations. *Journal of Applied Psychology*, 84(5), 721–736.
- Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, 19(2), 59–92. <https://doi.org/10.1177/1529100618808244>

- Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist*, 53(1), 5–26.
- Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. Springer.
- Hao, J., Liu, L., von Davier, A. A., Lederer, N., Zapata-Rivera, D., Jakl, P., & Bakkenson, M. (2017). EPCAL: ETS platform for collaborative assessment and learning. *ETS Research Report Series*, 2017(1), 1–14.
- Jeong, H., Hmelo-Silver, C. E., & Jo, K. (2019). Ten years of Computer-Supported Collaborative Learning: A meta-analysis of CSCL in STEM education during 2005–2014. *Educational Research Review*, 28, 100284. <https://doi.org/10.1016/j.edurev.2019.100284>
- Kerr, D., Andrews, J. J., & Mislevy, R. J. (2016). The in-task assessment framework for behavioral data. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (pp. 472–507). Wiley.
- Ko, M. M., Chandler, C., Rose, S., Cline, B., Watts, E., Reitman, J., Foltz, P. W., & D’Mello, S. K. (2016). AI Partners that Support Productive Uncertainty Within “Jigsaw” Activities During Small Group Collaborative Learning in Classrooms. In *International Conference on Artificial Intelligence in Education* (pp. 608–622). Cham: Springer Nature Switzerland.
- Manz, E., & Suárez, E. (2018). Supporting teachers to negotiate uncertainty for science, students, and teaching. *Science Education*, 102(4), 771–795.
- Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated Scoring of Complex Tasks in Computer-Based Testing* (pp. 15–47). Routledge.
- OECD. (2017). *PISA 2015 results: Collaborative problem solving*. OECD Publishing.
- Penuel, W. R., Philip, T. M., A. Chang, M., Sumner, T., & D’Mello, S. K. (2026). Responsible Innovation in Designing AI for Education: Attending to the “How,” the “for What,” the “for Whom,” and the “With Whom” in a Rapidly Growing Field. *Educational Researcher*, 55(2), 134–141. <https://doi.org/10.3102/0013189X251389879>
- Southwell, R., Pugh, S., Perkoff, M., Clevenger, C., Bush, J. B., Lieber, R., ... & D’Mello, S. K. (2022). Challenges and feasibility of automatic speech recognition for modeling collaborative discourse in classrooms. *Proceedings of the 15th International Educational Data Mining Conference (EDM 2022)* (pp. 302–315).
- Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research*, 69(1), 21–51.
- Sun, C., Shute, V., Stewart, A., Yonehiro, J., Duran, N., & D’Mello, S. K. (2020). Toward a generalized competency model of collaborative problem solving. *Computers & Education*, 143, 103672.
- Watkins, J., & Manz, E. (2022). Characterizing pedagogical decision points in sense-making conversations motivated by scientific uncertainty. *Science Education*, 106(6), 1408–1441.
- World Economic Forum. (2025). *The future of jobs report 2025*. <https://www.weforum.org/publications/the-future-of-jobs-report-2025/in-full/3-skills-outlook/>
- Xu, E., Wang, W., & Wang, Q. (2023). The effectiveness of collaborative problem solving in promoting students’ critical thinking: A meta-analysis based on empirical literature. *Humanities and Social Sciences Communications*, 10(1), 1–11.