

LERN

Learning Engineering Research Network

2026 CONVENING

From Insights to Implementation
Learning Engineering in Action

Convening Proceedings

2026

Proceedings of the Learning Engineering Research Network Convening (LERN 2026)

From Insights to Implementation, Learning Engineering in Action

Scotty D. Craig, Tracy Arner, & Danielle S. McNamara

learning engineering

learning engineering proceedings

Learning Engineering Process

Learning Engineering Research Network

LERN

6

“





DOI: 10.59668/2551

2026

🔍

Filter content...

Applied Learning Engineering Research & Testing (ALERT)

What Works When for Whom Under What Conditions: Learning Engineering as an Enabler of Component-based Research

Learning Engineering Body of Knowledge

A Tiered Framework for Educational Event Data Documentation: Synthesizing Principles and Addressing Gaps

IEEE Standards for Scaled Learning Engineering

Learning Engineering by Design: An Agentic AI Application for Rapid, Personalized Health & Safety Training in Disaster Response and Hazardous Environments

Scaling Interleaved Mathematics Practice

Teacher Education through Learning Engineering: An Action Research on Faculty Transformation

Understanding Instructor Perspectives and Course Challenges in FSE 100: A Learning Engineering Approach to Improving the Course

Human Centered Applications within Learning Engineering

A Learning Engineering Approach to Transforming Teacher Practice Through Co-Designing Science Curricular for Multilingual Learners

A Learning Strategy Analysis for Guiding the Creation of a Team Training Immersive Reality Environment

Advancing Usability of an Immersive Virtual Reality Team Training Environment within a Learning-Engineering Cycle

Case Study: Learning Economy Foundation Competency Graph

Co-Designing a Study Platform for ASU Graduate Students to Enhance Learning Productivity and Performance Through Connection

Co-Designing AI-Enabled Learning in Nursing Education: A Learning Engineering Approach Using Faculty and Student Insights

Developing a Model to Support Collaborative Engineering Projects: Student Uncertainty as a Productive Resource for Engineering (SUPER-E)

Developing Learning Strategy Heuristics for Active Mobile Learning Platforms

Human-Centered Learning Ecosystems: Reimagining Water Education for Real Estate Professionals

AI Systems Informed by Learning Engineering

A Predictive Learning Engineering Framework for Modeling Active Learning

Adaptive Multi-Modal Deepfake Detection for Safer Learning Environments

Agentic PAL: Designing Human-Empowered AI Partnerships for Early Childhood Mathematics Learning

An Application of Design-Based Implementation Research to Develop a Framework to Support a Community of AI-experience Creators

Automated Run-on Sentence Detection and Correction for Educational Writing

Automatic Identification and Evaluation of Revisions in Student Writing Using Large Language Models

Automatically Generating Interactive Learning Experiences with an LLM-Driven Agentic Pipeline

Bridging Human Intelligence Augmentation (IA) and Classroom Practices via GenAI in Learning Engineering

Charm-bots: The Impact of A.I.'s Sycophancy Language on User Trust

Coaching, Not Autocomplete: Early Evidence from ConnectInk's AI-Supported Personal Narrative Pilot

Comparing Epistemic Emotions and User Experience Across Two AI Instructional Designs in Biology Learning

Currents of Inquiry: Insights From Two Years of Real-World AI-Learner Water Conversations

Design and Pilot Evaluation of a Gamified Narrative Chatbot for STEM Education

Designing for Student Engagement with AI in Courseware: Lessons from Iterative Improvements to DOT in REAL CHEM

EdLight Research Portal: An Expert-Annotated Repository of Handwritten Math Student Work

Epistemic Cognition and Uncertainty Navigation with a Domain-Specific AI Chatbot in STEM Education

From Course Concept to Lecture Video: An AI-Powered System for Automated MOOC Development

From Measurement to Action: A Learning Engineering Approach to AI-Powered Assessment for Human Power Skills Development

Implementing Concept Instruction via MCP Server

L2-French Learners and Generative AI (GenAI): Challenges, Needs, and Design Guidelines

Learning-By-Explaining with Generative AI: A Pilot Implementation in Introductory Biology

LLM Safety in an Educational Context: A holistic approach

Media Mentor AI: How a SCAMPER-guided AI assistant is helping reimagine media literacy learning

MIRANDA: Real-Time Learning Analytics for Authentic Embedded Assessment

Multiple-Document Comprehension in High School Science: A Learning-Engineering Pilot Study

NLP Validation of Prompt Strategies for Theory-Aligned LLM-Generated Personalization

Open Repository for AI Models as Learning Engineering Components

Optimizing Language-Focused Writing Feedback from Large Language Models through Prompt Engineering

Quality Assessment Through Learning Engineering: An Evaluation Rubric of LLM-Generated Multiple-Choice Questions

Reasoning LLMs are Competent Courseware Reviewers

ReQUESTA: A Hybrid Agentic Framework for Generating Cognitively Diverse Multiple-Choice Questions

Social and Emotional Dimensions of Generative AI Use

Socio-Emotional Learning in AI K-12 Guidance and Policy Documents: A Gap Analysis

The Difficult Conversations Bot: Findings on Fostering Empathy and Reflective Communication

Among Faculty and Staff

The Writing Analytics Tool: A Learning Engineering Approach to Designing AI-Supported Writing Instruction

Towards Automated Detection of Struggling Student Programmers

User experience design of AI-assisted human-technology ecosystem for writing assessment

Using AI to Bridge Technology Gaps in Higher Education

Educational Technology Informed by Learning Engineering

CoreTrust as a Civic Learning Tool: Community Capital for Inclusive Digital Governance

Engineering a Semantic Topicality Instrument for Multiple-Choice Question Quality Control

Design and Development of Two Digital Solutions for Promoting Harmonious Relationships: Buddy Up and Squeez

From Immersion to Action: How VR Influences Behavioral Intentions Around Advanced Water Purification

From Immersion to Behavioral Intentions: Virtual Reality and Advanced Water Purification

How Students Really Use Courseware: Visualizing Student Pathways in Integrated Chemistry Courseware

REAL CHEM Action Research Through the LearnLab Summer School

SCALED: A Transferable Learning-Engineering Model for Large-enrollment STEM Courses

Terracotta: Lowering Barriers to Experimental Education Research

The Education Tree: A New Theoretical Model for P-20 Education and Development

The Impact on Cognition and Motivation Using Gaming, Simulation, and Visual Learning in Military Flight Training

The Promise of Scenario-Based Assessments for College Instruction

“Walk It Out”: An Embodied and Mobile AI Tutor for STEM Education

Impact of Varying the Playback Speeds of Educational Content on Learning and Engagement

Proceedings of the Learning Engineering Research Network Convening

From Insights to Implementation: Learning Engineering in Action

Scotty D. Craig, Tracy Arner, & Danielle S. McNamara

The Learning Engineering Research Network (LERN) Convening is an ongoing conference dedicated to fostering top-quality research in Learning Engineering. The conference offers an interdisciplinary outlet for research in cognitive science, educational science, and design sciences, with a focus on developing effective learning ecosystems.

The 2026 convening, themed “From Insights to Implementation: Learning Engineering in Action,” brings together researchers, practitioners, and innovators to advance the interdisciplinary field of Learning Engineering. Building on work from cognitive science, education, design sciences, and beyond, the convening emphasizes translating research insights into practical solutions that strengthen learning ecosystems.

Hosted by the ASU Learning Engineering Institute, this collaborative event features keynotes, panels, and research presentations that showcase how data, technology, and design can drive innovation in education. From exploring generative AI to advancing inclusive practices, the LERN Convening highlights actionable approaches to addressing critical challenges and promoting equity in learning.

Held from February 3-4 in 2026, LERN2026 brought together over xxx educators, researchers, technologists, and learning professionals from public, private, and academic sectors. The LERN proceedings includes 70 examples of research within the learning engineering field from 95 submissions. These papers span from core papers on the Learning Engineering Process (LEP), human-centered evaluations within LEP, educational technology systems informed by the LEP, and the largest group, Artificial Intelligence in Education.

LERN2026 Awards

Of these submissions, we awarded two best paper awards. The awards were selected based on recommendations from reviewers and approval by our award committee.

2026 LERN Convening Best Paper Award

Human-Centered Learning Ecosystems: Reimagining Water Education for Real Estate Professionals

Danielle Storey, Poorva Ketkar, Emma Noble, & Harry Cooper

Coaching, Not Autocomplete: Early Evidence from ConnectInk's AI-Supported Personal Narrative Pilot

Julio Intriago-Izquierdo & Rahul Patel

Learning Engineering Research Network Distinguished Fellows (2026)

At this convening, we established our LERN Distinguished Fellows. The hard work and dedication of our three inaugural fellows have provided guidance and wisdom, and made the field of learning engineering what it is today, while setting it up to continue into the future. Our awardees this year are Sae Schatz, Jim Goodell, and Chris Dede.

Sae Schatz

Sae Schatz has been a key contributor to the evolution of learning engineering, especially through her work at the intersection of learning science, data-informed design, and advanced training technologies. As the former director of the Advanced Distributed Learning (ADL) Initiative for the U.S. Department of Defense, she helped shape modern learning ecosystems—championing interoperability standards, learner-centric design, and evidence-based approaches to large-scale training. Her leadership has pushed the field toward more adaptive, data-rich, and connected learning environments that embody the core principles of learning engineering.

Jim Goodell

Jim Goodell has been one of the central architects of learning engineering as a formal discipline, especially through his leadership in defining its frameworks, standards, and professional practices. As editor of Learning Engineering Toolkit and a longtime leader within the IEEE Learning Technology Standards Committee and ICICLE, he has advanced the field's focus on evidence-based design and continuous improvement. His work has helped establish learning engineering as a rigorous, systems-oriented approach to designing.

Chris Dede

Chris Dede has been one of the most influential voices shaping the modern field of learning engineering. A longtime Harvard professor and pioneer in learning technologies, he helped define learning engineering as the systematic use of data, design, and iterative experimentation to improve learning at scale. His work spans immersive learning (AR/VR), digital ecosystems, and large-scale research–practice partnerships, and he co-edited Learning Engineering for Online Education, a foundational text that formalized the discipline's methods and principles.

Convening Program and Review Committee

Ruben Acuña, Arizona State University, School of Computing and Augmented Intelligence

Tracy Arner, Learning Engineering Institute, Arizona State University

Christine M. Covas-Smith, Ph.D., Air Education and Training Command

Dr. James Dunnigan, Mary Lou Fulton College

Ariah Elmore, iPerformX

Md Biplob Hosen, University of Maryland Baltimore County

Tanya Churaman, University of Missouri-Columbia; Pearson Education

Linh Huynh, Learning Engineering Institute

Mina C. Johnson-Glenberg, Arizona State University

Heather C. Lum, Arizona State University

Mei Mei Li, Learning Engineering Institute, Arizona State University

Li (Lee) Liang, EvalMate.AI and Lee Vision

Janice Mak, Learning Engineering Institute, Arizona State University

Bryan Matlen, WestEd

Sarah Martin, Arizona State University

Kathryn S. McCarthy, Department of Learning Sciences, Georgia State University

Danielle S. McNamara, Learning Engineering Institute, Arizona State University

Chenyou Nicole Wu, NA

Catheryn Reardon, Learning Engineering Institute, Arizona State University

Henry Ryng, inXsol LLC

Anqi Shao, Arizona State University

Abigail Stein, Carnegie Mellon University

Danielle Storey, Arizona State University, Arizona Water Innovation Initiative

Yu Tian, Learning Engineering Institute, Arizona State University

Vipin Verma, Learning Engineering Institute, Arizona State University

Dr. Wendy Walsh, SL, DAFC, Chief Learning Officer HQ Air Education & Training Command

Yiqiao Xu, MetLife, Inc.

LERN 2026 Presentations

Keynote Speaker Sae Schatz, Ph.D.

CEO and Founder, the Knowledge Forge LLC



Thriving in the Age of Acceleration

We stand at a pivotal moment in the “Fourth Industrial Revolution,” our modern age, which isn’t simply defined by technological breakthroughs but also by the profound societal ripples they create. This keynote explores three defining characteristics reshaping how we work, live, and learn, and it provides a forward-looking framework for how Learning Engineering principles can help us build systems that match the pace of change.

About Sae

Sae works at the intersection of cognition, technology, and data. She formerly directed the Advanced Distributed Learning (ADL) Initiative, a government program for researching learning technologies, and before joining the civil service, Sae worked as an applied human–systems scientist in both business and academia, including as an assistant professor with the University of Central Florida’s Institute for Simulation and Training. Sae is a prolific writer and professional presenter; for example, she recently released [Engines of Engagement: A Curious Book About Generative AI](#) (2023) and contributed to the National Academies report on [Adult Learning in the Military Context](#) (2024).

Learning Engineering Research Network Convening Schedule

Welcome, Awards & Panel	Taking Human-Centered Learning Innovation to Scale	Moderator: Danielle McNamara, Executive Director of the ASU Learning Engineering Institute Panel members Steve Ritter, Founder and Chief Scientist of Carnegie Learning Alyssa Friend Wise, Professor of Technology and Education at Vanderbilt University and Director of the LIVE Learning Innovation Incubator Cristina Heffernan, Co-Executive Director and Co-Founder of The ASSISTments Foundation Ben Motz, Assistant Professor at Indiana University Bloomington

Best Paper

Human-Centered Learning Ecosystems: Reimagining Water Education for Real Estate Professionals	Danielle Storey
Coaching, Not Autocomplete: Early Evidence from ConnectInk's AI-Supported Personal Narrative Pilot	Julio Intriago-Izquierdo

Designing AI for K–12

Agentic PAL: Designing Human-Empowered AI Partnerships for Early Childhood Mathematics Learning	Anastasia Betts
Multiple-Document Comprehension in High School Science: A Learning-Engineering Pilot Study	Andrew Potter
Socio-Emotional Learning in AI K-12 Guidance and Policy Documents: A Gap Analysis	Emmanuel Adeloju

Applications of Learning Engineering

Teacher Education through Learning Engineering: An Action Research on Faculty Transformation	Kürşat Çağıltay

	eBuku: AI-Driven Mentorless Learning for STEM Education in Humanitarian Contexts of the DRC	Narcisse Mbunzama
	The Writing Analytics Tool: A Learning Engineering Approach to Designing AI-Supported Writing Instruction	Danielle McNamara & Andrew Potter
AI In Science and Education		
	Comparing Epistemic Emotions and User Experience Across Two AI Instructional Designs in Biology Learning	Yiwen Li
	REAL CHEM Action Research Through the LearnLab Summer School	Bryan Henderson
	Designing for Student Engagement with AI in Courseware: Lessons from Iterative Improvements to DOT in REAL CHEM	Kimberly Larson
Learning Engineering and Learning Technology 1		
	From Course Concept to Lecture Video: An AI-Powered System for Automated MOOC Development	Hua Wei
	MIRANDA: Real-Time Learning Analytics for Authentic Embedded Assessment	Elina Ollila
	The Impact on Cognition and Motivation Using Gaming, Simulation, and Visual Learning in Military Flight Training	Ariah Elmore
Debate	From Insights to Implementation: Learning Engineering in Action	Moderator: Tracy Arner, Associate Director of the Learning Engineering Institute Debate Participants Nia Nixon, Associate Professor in School of Education at the University of California, Irvine Stephen Fancsali, Vice President of Data Science at Carnegie Learning Kathryn McCarthy, Associate Professor of Educational Psychology at Georgia State University

René Kizilcec, Associate Professor of
Information Science at Cornell University

Learning Engineering Distinguished Fellows		
	What Works When for Whom Under What Conditions: Learning Engineering as an Enabler of Component-based Research	Chris Dede
	Learning Engineering Body of Knowledge	Jim Goodell
Education, Validity, & AI		
	The Education Tree: A New Theoretical Model for P-20 Education and Development	Maxwell Goshert
	Implementing Concept Instruction via MCP Server	Thor A. Anderson
	NLP Validation of Prompt Strategies for Theory-Aligned LLM-Generated Personalization	Linh Huynh
Learning Engineering and Learning Technology 2		
	"Walk It Out": An Embodied and Mobile AI Tutor for STEM Education	Mina Johnson-Glenberg
	A Tiered Framework for Educational Event Data Documentation: Synthesizing Principles and Addressing Gaps	Xin Wei
	Learning Engineering by Design: An Agentic AI Application for Rapid, Personalized Health & Safety Training in Disaster Response and Hazardous Environments	Henry Rying
AI & Pedagogy		
	ReQUESTA: A Hybrid Agentic Framework for Generating Cognitively Diverse Multiple-Choice Questions	Yu Tian
	Towards Automated Detection of Struggling Student Programmers	Sanjita Patwardhan
	The Promise of Scenario-Based Assessments for College Instruction	Jonathan Cohen

Human Centered
Applications with
Learning Engineering

User experience design of AI-assisted human-technology ecosystem for writing assessment	Li (Lee) Liang
A Learning Engineering Approach to Transforming Teacher Practice Through Co-Designing Science Curricular for Multilingual Learners	Yernat Mnuar, Jie Zhang, Iftekharul Chowdhury
Currents of Inquiry: Insights From Two Years of Real-World AI-Learner Water Conversations	Stephen Carradini

Poster Session

Poster Title	Presenter
Adaptive Multi-Modal Deepfake Detection for Safer Learning Environments	Syeda Samira Sama
An Application of Design-Based Implementation Research to Develop a Framework to Support a Community of AI-experience Creators	Janice Mak
Automated Run-on Sentence Detection and Correction for Educational Writing	Shubham Chakraborty
Automatic Identification and Evaluation of Revisions in Student Writing Using Large Language Models	Yu Tian
Automatically Generating Interactive Learning Experience With an LLM-driven Agentic Pipeline	Felix Gröner
Charm-bots: The Impact of A.I.'s Flattering Language on User Trust	Heather C. Lum
Engineering a Semantic Topicality Instrument for Multiple-Choice Question Quality Control	Michelle Banawan
From Measurement to Action: A Learning Engineering Approach to AI-Powered Assessment for Human Power Skills Development	Yigal Rosen
Integrating AI into Data Sensemaking: Teachers' Learning and Pedagogical Reflections	Emmanuel Adeloju
L2-French Learners and Generative AI (GenAI): Challenges, Needs, and Design Guidelines	Jiachen Gong
Learning-By-Explaining with Generative AI: A Pilot Implementation in Introductory Biology	Kathryn S. McCarthy
Leveraging Agentic AI for Human-Centered XR Learning Content Standards	Myranda Pina

LLM Safety in an Educational Context: A holistic approach	Bernice Yeo
Media Mentor AI: How a SCAMPER-guided AI assistant is helping reimagine media literacy learning	Karina Luna
Optimizing Language-Focused Writing Feedback from Large Language Models through Prompt Engineering	Yu Tian
QA Automation of Canvas Courses	Natalia Echeverry
Quality Assessment Through Learning Engineering: An Evaluation Rubric of LLM-Generated Multiple-Choice Questions	Katerina Christhlf
Reasoning LLMs are Competent Courseware Reviewers	Ryan Dwyer
Social and Emotional Dimensions of Generative AI Use	Rebekah Jongewaard
The Difficult Conversations Bot: Findings on Fostering Empathy and Reflective Communication Among Faculty and Staff	Catheryn Reardon
The Writing Analytics Tool: A Learning Engineering Approach to Designing AI-Supported Writing Instruction	Danielle McNamara
Epistemic Cognition and Uncertainty Navigation with a Domain-Specific AI Chatbot in STEM Education	Yiwen Li
Vibe Coding: How Junior and Senior Developers Use GenAI's Newest Tool	Selena Evans
A Learning Strategy Analysis for Guiding the Creation of a Team Training Immersive Reality Environment	Robert F. Siegle
Advancing Usability of a VR Team Training Environment within a Learning-Engineering Cycle	Parkhi Malhotra
Co-Designing a Study Platform for ASU Graduate Students to Enhance Learning Productivity and Performance Through Connection	Fiodesy Gemilang Putri
Co-Designing AI-Enabled Learning in Nursing Education: A Learning Engineering Approach Using Faculty and Student Insights	Mamta Shah
Design and Pilot Evaluation of a Gamified Narrative Chatbot for STEM Education	maria goldshtein
Developing Learning Strategy Heuristics for Active Mobile Learning Platforms	Ishrat Ahmed
How Students Really Use Courseware: Visualizing Student Pathways in Integrated Chemistry Courseware	David J. Yaron
Understanding Instructor Perspectives and Course Challenges in FSE 100: A Learning Engineering Approach to Improving the Course	Whitney Hansberry

Case Study: Learning Economy Foundation Competency Graph	Jim Goodell
IEEE Standards for Scaled Learning Engineering	Jim Goodell
Open Repository for AI Models as Learning Engineering Components	Jim Goodell
Bridging Human Intelligence Augmentation (IA) and Classroom Practices via GenAI in Learning Engineering	Li (Lee) Liang
A Predictive Learning Engineering Framework for Modeling Active Learning	Michelle Banawan
CoreTrust as a Civic Learning Tool: Community Capital for Inclusive Digital Governance	Mark Roseland
Design and Development of Two Digital Solutions for Promoting Harmonious Relationships: Buddy Up and Squeez	Kimberly R. M. Osborne
EdLight Research Portal: An Expert-Annotated Repository of Handwritten Math Student Work	Michel Meneses
From Immersion to Action: How VR Influences Behavioral Intentions Around Advanced Water Purification	Ketevan Chachkhiani
Terracotta: Lowering Barriers to Experimental Education Research	Benjamin Motz
Developing a Model to Support Collaborative Engineering Projects	Yu Ye
Impact of Varying the Playback Speeds of Educational Content on Learning and Engagement	Tyree Cowell
NSF: IGE: Transforming Master's-Level Engineering Education through Industry Partnerships, Principled Engineering, and Experiential Learning	Samantha Brunhaver
SCALED: A Transferable Learning-Engineering Model for Large-enrollment STEM Courses	Medha Dalal
Scaling Interleaved Mathematics Practice	Bryan Matlen
Curiosity Games: Project Mars: An AI-Supported Education Recommendation System	Caila DeAbreu

Acknowledgements

We would like to acknowledge our student volunteers for their help with the conference.

Parkhi Malhotra

Whitney Hansberry

Mozhgan Ansari

Fiodesy Gemilang Putri

Emma Streberger

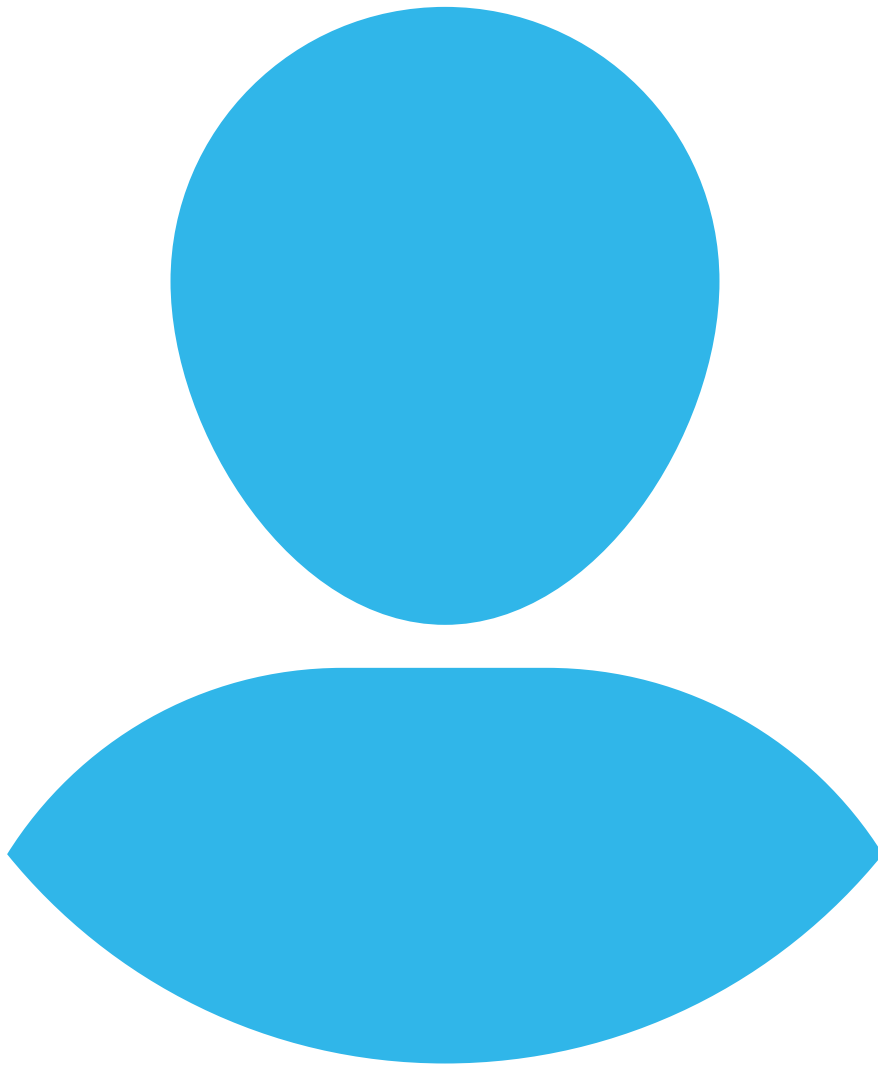
Emily Machniak

Jessica Tan

Emmanuel Adeloju

Rebekah Jongewaard

Sathkeerthi S V

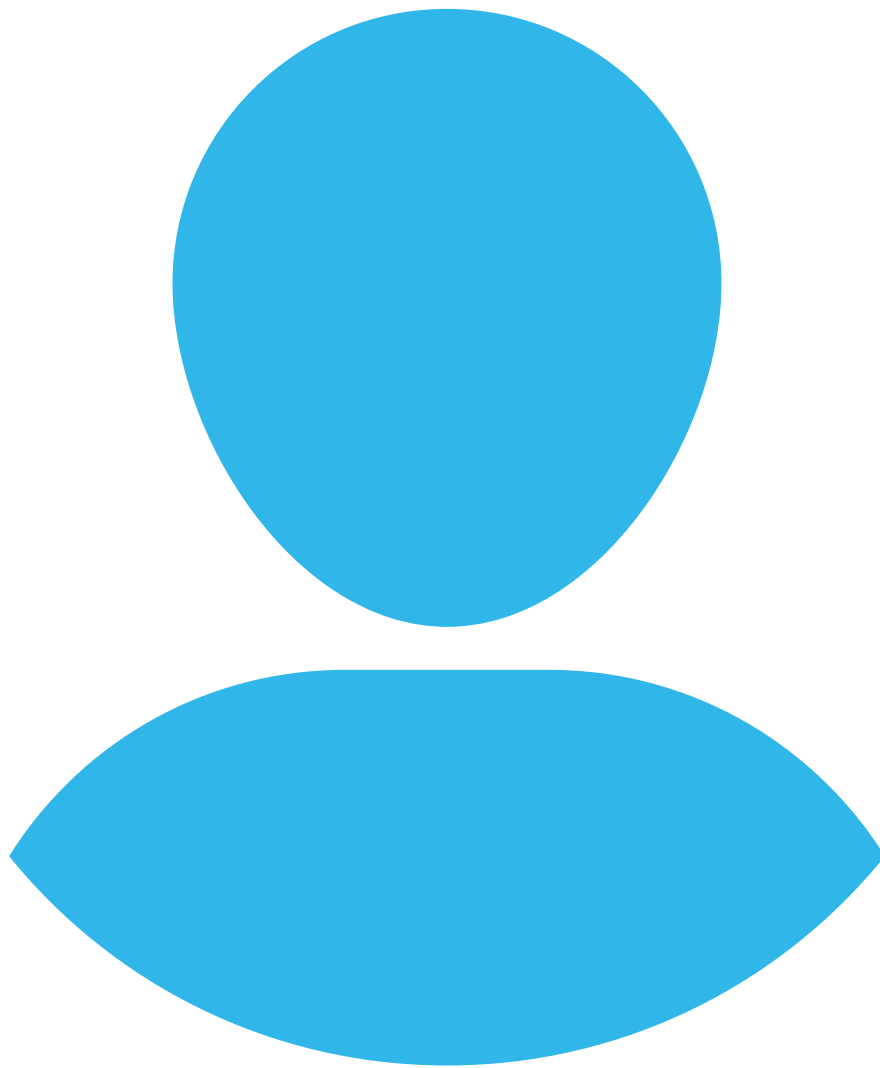


Scotty D. Craig

Arizona State University

Scotty D. Craig is an associate professor of human systems engineering within The Polytechnic School of the Ira A. Fulton Schools of Engineering at Arizona State University. Dr. Craig serves as the ASU Learning Engineering Institute's Director, Research and Evaluation and as Director of the ASU Advanced Distributed Learning Initiative Partnership Lab. Dr. Craig is a learning engineer with expertise in cognitive science, design science, and the science of learning (specifically learning technology).

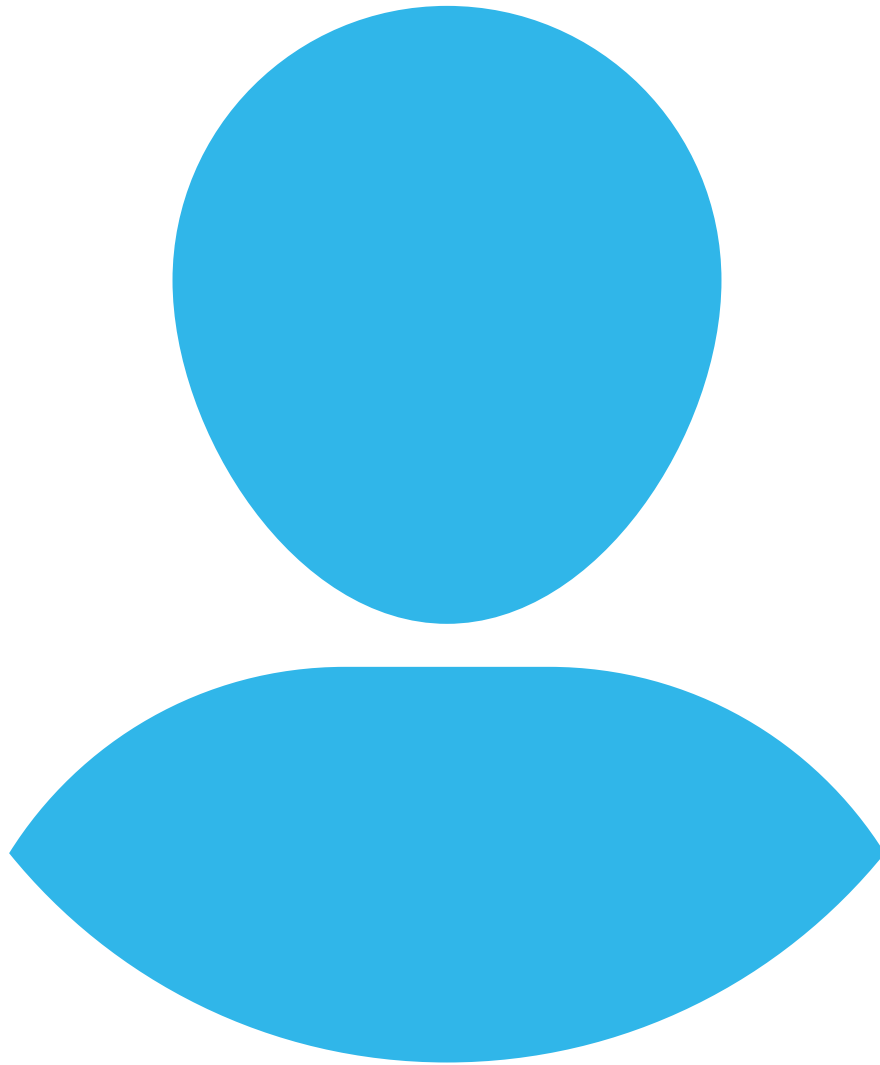
He has contributions at the intersection of psychology, education, and technology within the areas of multimedia learning, virtual environments, as well as development and evaluation of learning technology within laboratory and applied settings. Dr. Craig is currently the graduate chair for four areas: Human Systems Engineering Program (MS & PhD), User Experience (MS), Data Science and Analytics in Engineering- Human-Centered Applications (MS) and the new Graduate Certificate for Learning Engineering.



Tracy Arner

Arizona State University

Tracy Arner is the associate director for the Learning Engineering Institute at Arizona State University. She received her PhD in educational psychology from Kent State University in 2020. Her research interests include reading comprehension, instructional design, technology-enhanced instruction, game-based learning, and misconceptions.



Danielle S. McNamara



Applied Learning Engineering Research & Testing (ALERT)

[What Works When for Whom Under What Conditions: Learning Engineering as an Enabler of Component-based Research](#)

[Learning Engineering Body of Knowledge](#)

[A Tiered Framework for Educational Event Data Documentation: Synthesizing Principles and Addressing Gaps](#)

[IEEE Standards for Scaled Learning Engineering](#)

[Learning Engineering by Design: An Agentic AI Application for Rapid, Personalized Health & Safety Training in Disaster Response and Hazardous Environments](#)

[Scaling Interleaved Mathematics Practice](#)

[Teacher Education through Learning Engineering: An Action Research on Faculty Transformation](#)

[Understanding Instructor Perspectives and Course Challenges in FSE 100: A Learning Engineering Approach to Improving the Course](#)



What Works When for Whom Under What Conditions: Learning Engineering as an Enabler of Component-based Research

Chris Dede

Artificial Intelligence

Component-based Research

learning engineering

Scaling through Personalization

Component-based research (CBR) is a methodological strategy centered on studying the features and processes of innovations that contribute to desired outcomes for specific types of learner populations, conditions, and contexts. CBR focuses on precisely defined elements of overall innovations, called components, that can be studied individually or in clusters across multiple implementation sites. Compatible language in describing data, specificity in delineating the components of innovations, and processing data combined across studies in a principled manner are all attributes of learning engineering (LE) vital for CBR. Measurements and analyses can then determine not only if the innovation was effective, but also which parts of the innovation are correlated with outcomes and situations. These methodological advances depend on developing data infrastructures, enabled by artificial intelligence and nested LE, that empower component-based design, implementation, and evaluation. This is essential for

personalized learning as well as for achieving scale through principled adaptation to local conditions for success based on the relative contributions of each component.

Introduction: An Aspirational Vision for Educational Research

The ability of educational research to answer the question, “What parts of an educational innovation work for whom and under what conditions?” would be a major advance in achieving educational impact (Means, 2022). Instructional designers could then evolve effective innovations targeted to specific settings and learners. Teachers could personalize interventions to individual learners, and administrators could use evidence to determine which innovations should be implemented in their context. Technical assistance providers could aid decision makers in applying the most recent research insights to their unique setting, and policy makers could develop evidence-based targeted strategies for support, regulation, and continuous improvement. Funders could allocate their resources based on proven strategies for personalization directed towards equity and excellence.

Such a systemic and aligned flow of advances can be achieved via component-based research (CBR). This methodological strategy centers on studying the features and processes of innovations that most contribute to desired outcomes for specific types of learner populations, conditions, and contexts (Century, 2022). In contrast to conventional educational research methods, CBR is based on precisely defined elements of overall innovations, called components, that can be studied individually or in clusters across implementation sites. In its research, CBR requires describing specific attributes of both settings (characteristics, conditions) and innovation participants (sociodemographic factors, self-perceptions, life experiences) that influence the outcomes of the intervention. Another condition for success of CBR is the detailed specification of desired outcomes (Century, Dede, Taylor, Brooks, Han, Scher, & Tutwiler, 2024).

Learning engineering (LE) is essential for achieving these dimensions of successful CBR. Compatible language in describing data, specificity in delineating the components of innovations, and combining data across studies in a principled manner are all attributes of LE vital for CBR. Using the full range of LE principles, attributes, and strategies, the value of educational research can go deeper in personalized interventions than determining the average effectiveness of an overall intervention across populations, conditions, and contexts (Dede, 2019). Measurements and analyses can determine not only if the innovation was effective, but also what parts of the innovation are correlated with outcomes and situations, as delineated in the related methodology of design-based implementation studies (Sabelli & Dede, 2013). LE provides an infrastructure of strategies, measures, and analytics; technical capabilities for data storage, use, and management; and evolving approaches based on advances in data science and artificial intelligence. All this enables accumulating component-based findings across studies and innovations to develop insights about what works for whom under what conditions.

This paper considers how LE can enable the efficient flow of CBR to scale innovations from effective use of data in design to implementation studies across different contexts.

Learning Engineering Databases for Component-based Research

Examples from engineering physical artifacts can illustrate the types of component taxonomies and data infrastructures needed for CBR in education. To meet the challenge of building a bridge over a body of water, engineers draw on a well-specified set of components that characterize various types of bridges (e.g., arch, beam, cantilever, suspension). Each type has its strengths and limits depending on the purpose of the bridge (i.e., heavy highway traffic, footbridges), the materials and budget available, and the attributes of the physical setting where the bridge is to be placed (e.g., water shear, substrate under foundations, length, maximum stress, minimizing harmonic vibration). Given the precise specification of these characteristics and components, engineers are able to conduct CBR to determine the conditions under which a suspension bridge will be successful at a particular site and for a designated purpose.

Current educational research struggles to do this type of CBR. Terms like “collaborative learning” are ambiguously defined and often underspecified in scholarly publications about the implementation studied. Further, going from the learning sciences to detailed design protocols is like going from the physics and chemistry involved in a bridge straight to construction site blueprints; intermediate LE heuristics and strategies (e.g., spaced reinforcement is effective in aiding retention) are essential (Goodell, Kolodner, & Kestner, 2023). Century et al (2024) provide a detailed delineation of the advances needed to fully apply learning engineering to CBR.

An Example of Applying LE to CBR: The Architecture for AI-Augmented Learning (A4L)

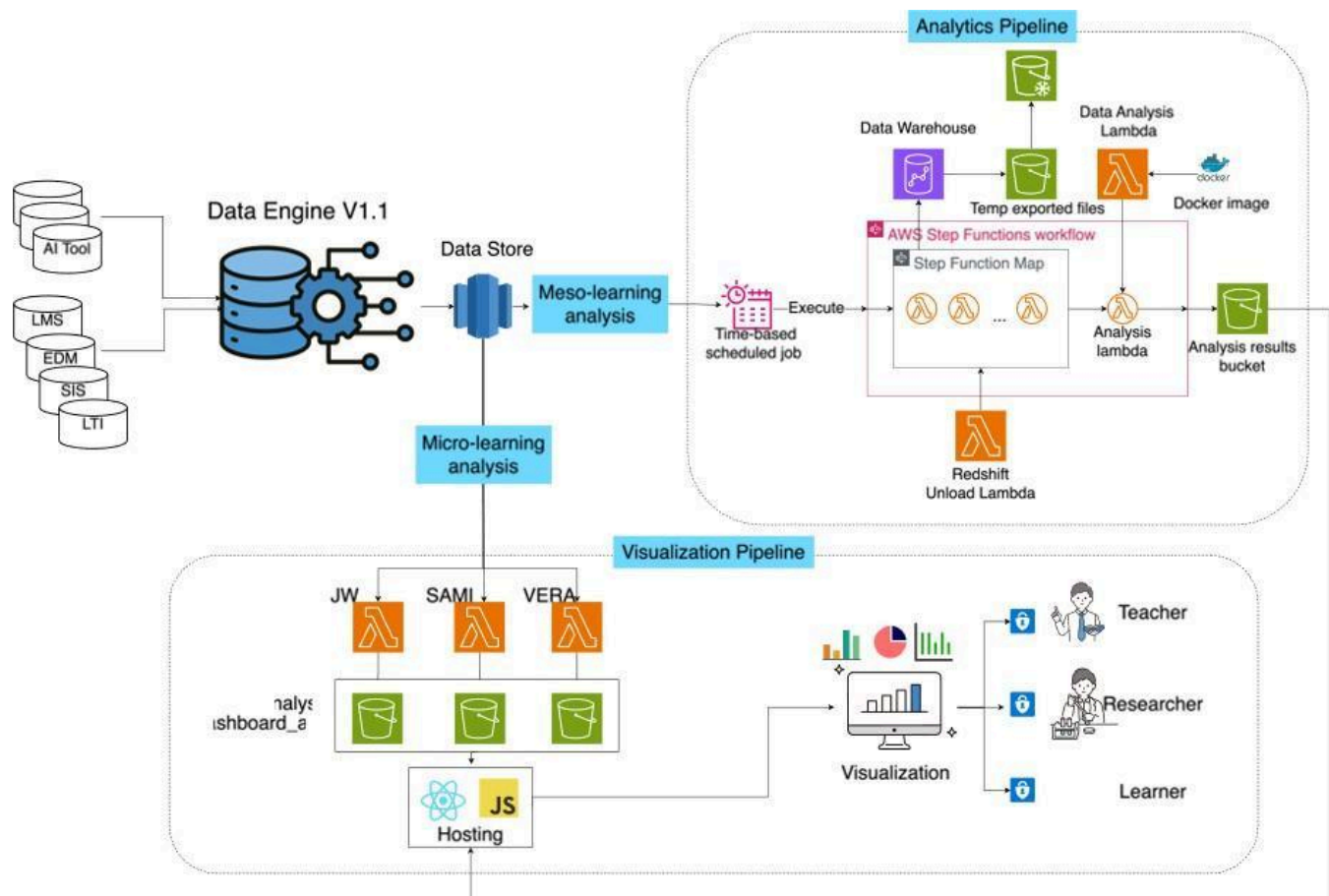
Educational research teams can use LE to design and develop data infrastructures that empower CBR. As an illustration, the National AI Institute for Adult Learning and Online Education (AI-ALOE) is developing a data architecture called A4L (Goel, Thajchayapong, Nandan, Sikka, & Rugaber, 2025). Across implementation sites, this pipeline can collect, combine, and analyze learning data at three levels:

1. Micro-learning data from individual learning episodes (e.g., clickstream, engagement metrics, cognitive patterns)
2. Meso-learning data over time (e.g., within-course student performance trends; longitudinal data about collaborative discussions and group work, learning-management-system [LMS] quizzes, feedback)
3. Macro-learning data about longer-term educational progress (e.g., cross-course performance, multi-source integration across LMS and AI tools)

1Edtech, which is a standards-setting association, is an important partner in enabling the interoperability of the various systems that collect these types of data.

Figure. 1.

Architecture for the A4L Data Pipeline



Constructing this data architecture involved a nested LE cycle. At each stage, designers made decisions based on supports similar to a LEED tracker (Totino & Kessler, 2024). As an example, the Visualization Pipeline translates complex data outputs into intuitive, role-specific dashboards (Thajchayapong & Goel, 2025). These dashboards are accessible to instructors, learners, and researchers, and they serve as a key interface for human-AI collaboration in the classroom. Drawing on principles from the learning sciences and cognitive science (Craig et al, 2025), educators use these visualizations to inform instructional design and pedagogical decisions, and learners use customized visualizations to chart their progress and self-regulate their learning processes. The dashboard visualizations also serve as a crucial way of sustaining an ongoing feedback loop between students and teachers, enabling continuous improvement.

The LE strategy for combining various kinds of data collected across time and location can convey insights about various types of components. As an illustration, AI-ALOE uses the Community of Inquiry (COI) framework for attaining motivation and effectiveness in online and blended education (Vaughan, Dell, Cleveland-Innes, & Garrison, 2023). The three components in COI are cognitive presence, teaching presence, and social presence; each is tightly defined and has a standardized assessment. Using A4L, AI-ALOE can do CBR to determine the effect of each presence-type across learning experiences and AI tools. Early-stage research has demonstrated the technical and practical feasibility of this approach (Thajchayapong & Goel, 2025).

Achieving Personalization at Scale through CBR

The aspirational vision of CBR data pipelines such as A4L is to enable powerful strategies for personalization in learning based on understanding which combinations of components work best for whom, under what conditions. The use of generative AI is enabling new opportunities for achieving this goal. As an illustration, the analytics underlying CBR are enriched by natural

language processing, and the emergence of AI-enhanced tools can enable digital communities of practice in which AI-assistants can share information and insights about individual learners across disparate learning experiences and extended timeframes.

Strategies for personalization are crucial to achieving scale for any learning innovation. Coburn (2003) defined scaling a successful innovation as encompassing four interrelated dimensions. These are 1) depth of changes to classroom practice and expectations; 2) sustainability of maintaining those changes over time; 3) spread to other practice contexts; and 4) shift in ownership of the reforms to practitioners. Dede (2006) extended Coburn's framework by adding a fifth dimension: the evolution that develops when users adapt the innovation in ways that influence and reshape the thinking of its designers, leading to a revised core model. By fostering evolution, designers can build practitioner capabilities for effective use, can highlight the role of research as a key component of implementation, and can serve as intermediaries between more general research knowledge and localized practices (Damschroder, Reardon, Widerquist, & Lowery, 2022).

Clarke and Dede (2009) applied this framework for scale to the implementation of a middle school science curriculum across a wide range of sites. The initial stage of the scaling strategy is Depth: Understanding why an innovation works well requires discovering the role of each component in contributing to its effectiveness. Then it is important to establish, through CBR based on LE, the changes in effectiveness that occur if some of these components are weakened or omitted. This frequently happens in adapting for scale, as few contexts will have all the conditions for success that nurtured the initial implementation of the innovation.

Two of the five interrelated dimensions of this scaling framework involve adaptation to personalize the innovation to a local set of conditions. The first is Sustainability: If adopters find that they lack some of the conditions for success in the original program, with insights from CBR they can develop variations of the innovations that better fit their own situation and are sustainable. Like hybrids in agriculture, the adapted innovation may not be as effective as the original version but still can yield valuable outcomes. The second is Spread: It may be necessary—and desirable—to modify a program to reduce the level of resources needed while retaining much of its effectiveness. For example, a highly effective innovation may scale best when a somewhat less powerful but still effective version involves a more affordable level of professional development. Findings from knowledge diffusion support this piloting strategy, documenting the importance of easily implemented trial-versions of an innovation (Dearing et al, 2015).

While this discussion implies a sense of progression, the five dimensions do not articulate a linear movement through phases, since the scaling strategies are interrelated in complex ways. Instead, the framework delineates multiple processes an implementation strategy can utilize in developing sustainable scale (Dymnicki, Trivits, Hoffman, & Osher, 2020). When CBR is achieved through LE, then leaders can make educational decisions based on predictions of how well an innovation will function for various types of learners given the local presence or absence of various conditions for success. Also, CBR studies of a widely scaled innovation can yield insights about the function and impact of particular components, which can inform a broad range of future designs.

Conclusion

The microscope and telescope each enabled collecting and analyzing data that had always been there to develop new findings, such as bacterial diseases in wellness and the role of dark matter in the universe. Through AI, LE and CBR, we now have the equivalent of the microscope and telescope for learning, and exciting new understandings are likely to emerge (Dede & Lidwell, 2023). For example, randomized clinical trials using research methods that measure outcomes as group averages often conclude that an intervention makes no significant difference. With CBR and LE, even if the intervention as a whole has little impact on the group as a whole, researchers can identify components of the intervention that may have a significant effect on subgroups. This enables new pathways towards design improvement and personalization.

Acknowledgments

This work was supported by US National Science grants #2247790 and

#2112532 to the National AI Institute for Adult Learning and Online Education (aialoe.org), as well as US National Science Foundation grant #2246543 to the University of Chicago.

References

- Century, J. (2022). The Case for component-based research. <https://datascience.uchicago.edu/wp-content/uploads/2023/03/The-Case-for-Component-based-Research.pdf>
- Century, J., Dede, C., Taylor, J., Brooks, J., Han, D., Scher, L., & Tutwiler, S. (2024). Component-based research in education: Emerging ideas, possibilities, and next steps. University of Chicago: NSF-funded Working Meeting on Component-based Research. <https://osf.io/preprints/edaxiv/vs3qw>
- Clarke, J., & Dede, C. (2009). Design for scalability: A case study of the River City curriculum. *Journal of Science Education and Technology* 18(4), 353-365.
- Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher*, 32(6), 3-12. <https://doi.org/10.3102/0013189X032006003>
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a nested learning engineering methodology to develop a team dynamic measurement framework for a virtual training environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). <https://doi.org/10.59668/2109.21735>
- Damschroder, L. J., Reardon, C. M., Widerquist, M. A. O., & Lowery, J. (2022). The updated Consolidated Framework for Implementation Research based on user feedback. *Implementation Science*, 17(1), 75. <https://rdcu.be/dHAEq>
- Dearing, J.W., Dede, C., Boisvert, D., Carrese, J., Clement, L., Craft, E., Gardner, P, Hyder, J., Johnson, E., McNeel, D., Phiri, J., & Pleil, M. (2015). How educational innovators apply diffusion and scale concepts. In Chee-Kit Looi & Laik-Woon Teh, *Sustaining and Scaling Educational Innovations*, pp. 81-104. New York: Springer.
- Dede, C. (2006). Scaling up: Evolving innovations beyond ideal settings to challenging contexts of practice. In R.K. Sawyer (Ed.), *Cambridge Handbook of the Learning Sciences*, pp. 551-566. Cambridge, England: Cambridge University Press.
- Dede, C. (2019). Improving efficiency and effectiveness through learning engineering. In C. Dede, J. Richards, & B. Saxberg, (Eds.). (2019). *Learning engineering for online education: Theoretical contexts and design-based examples*, pp.1-14. New York: Routledge.
- Dede, C., & Lidwell, W. (2023). Developing a next-generation model for massive digital learning. *Education Sciences* 13(8), 845-854. <https://doi.org/10.3390/educsci13080845>
- Dymnicki, A., Trivits, L., Hoffman, C., & Osher, D. (2020). Advancing the use of core components of effective programs: Suggestions for researchers publishing evaluation results. Office of Assistant Secretary for Planning and Evaluation, US Department of Health and Human Services. https://youth.gov/sites/default/files/ASPE-Brief_Core-Components.pdf
- Goel, A., Thajchayapong, P., Nandan, V., Sikka, H., & Rugaber, S. (2025). A4L: An architecture for AI-augmented learning. *Computers and Society*. <https://doi.org/10.48550/arXiv.2505.06314>

- Goodell J., Kolodner, J., & Kessler, A. (2023). Tools from the learning sciences. In J. Goodell & J. Kolodner (Eds.), *Learning Engineering Toolkit*, pp. 243-253. New York, NY: Routledge.
- Means, B. (2022). Making insights from educational psychology and educational technology research more useful for practice. *Educational Psychologist*, 57(3), 226–230. <https://doi.org/10.1080/00461520.2022.2061974>
- Sabelli, N., & Dede, C. (2013). Empowering design-based implementation research: The need for infrastructure. In B. J. Fishman, W.R. Penuel, A-R Allen, & B.H. Cheng (Eds.), *Design-based implementation research: Theories, methods, and exemplars* (National Society for the Study of Education, Volume 112, Issue 2), pp. 464-480. NY, NY: Teachers College, Columbia.
- Thajchayapong, P., & Goel, A. K. (2025). Personalized learning through AI-driven data Pipeline. *Proceedings of the AAAI Symposium Series*, 5(1), 111-114. <https://doi.org/10.1609/aaaiss.v5i1.35572>
- Totino, L., & Kessler, A. (2024). “Why did we do that?” A systematic approach to tracking decisions in the design and iteration of learning experiences. *The Journal of Applied Instructional Design*, 13(2) <https://doi.org/10.59668/1269.15630>
- Vaughan, N.D., Dell, D., Cleveland-Innes, M., & Garrison, R.D. (2023). *Principles of blended learning: Share metacognition and communities of inquiry*. Toronto, CA: AU Press.



Learning Engineering Body of Knowledge

Jim Goodell

learning engineering

Learning Engineering Body of Knowledge

A Guide to the Learning Engineering Body of Knowledge (LEBOK Guide), soft-released in December 2025, will be formally released at LERN 2026. This open-source resource does for learning engineering what the Software Engineering Body of Knowledge (SWEBOK Guide) does for software. SWEBOK addresses principles and practices of designing, developing, testing, and maintaining software systems. LEBOK is a guide to the principles and practices of learning engineering. The Guide is released as a Wiki version (lebok.wiki), a PDF, and a machine-readable implementation in IEEE Sharable Competency Definition (IEEE 1484.20.3-2023) format used to link from learning resources, learning event metadata, and digital credentials. The wiki version of the guide is intended to be a platform for iterative community development and vetting, leading to future community-authorized releases. The Guide is currently organized into 12 knowledge areas (KAs) followed by appendices. Within each knowledge area are topics and subtopics.

Introduction

Learning engineering is a multidisciplinary practice and iterative process that applies the learning sciences, human-centered design, engineering principles and data informed decision making to support learners and their development. As the field has matured, a need emerged for a consistent, global view of the profession's core competencies. The Guide to the Learning Engineering Body of Knowledge (LEBOK) was developed to provide an organizing framework for these competencies, recognizing the diverse disciplines that contribute to effective learning engineering.

Other professional domains such as software engineering and project management have authoritative body-of-knowledge guides to help frame the scope and breadth of those professions. These other guides served as a model for LEBOK. And as learning engineering is a multidisciplinary practice that often requires expertise in software and project management, this Guide recognizes the authority of other body-of-knowledge guides including:

The Guide to the Software Engineering Body of Knowledge (SWEBOK) is rooted in work by the Software Engineering Institute (SEI) at Carnegie Mellon University. The SEI published its own technical report in 1999 titled "A Software Engineering Body of Knowledge Version 1.0" (CMU/SEI-99-TR-004). The SEI's work served as a foundational input for what eventually became the official SWEBOK. It was originally developed as a joint initiative between the IEEE Computer Society and the Association for Computing Machinery (ACM) to achieve professional consensus on the core competencies of software engineering. The first official release, SWEBOK 2004, was later adopted as an ISO technical report (ISO/IEC TR 19759:2005). IEEE continued its stewardship after ACM withdrew from the project, releasing Version 3.0 in 2014 to expand the initial 10 Knowledge Areas (KAs) to 15. Most recently, the 2024 publication of Version 4.0 has modernized the guide further, integrating contemporary practices like Agile, DevOps, and Artificial Intelligence to ensure the body of knowledge remains relevant.

A Guide to the Project Management Body of Knowledge (PMBOK® Guide) for the project management profession is published by the Project Management Institute (PMI). Since its first official edition in 1996, the guide has evolved from a collection of generally recognized good practices into a comprehensive framework used by millions of practitioners worldwide. With the release of the Seventh Edition (2021) and the recent Eighth Edition (2025), the guide has shifted from a process-based approach toward a Principle-Based model. This modern version moves away from rigid "recipes" and instead focuses on 12 Principles of project management and 8 Performance Domains, emphasizing value delivery, adaptability, and the integration of Agile and Hybrid methodologies to suit the complexities of the modern digital landscape.

Guide to the Learning Engineering Body of Knowledge (LEBOK) draws from years of work by the IEEE International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) in defining the practice and process of learning engineering. It also draws from the corpus of knowledge produced by the ICICLE and learning engineering research communities.

Methodology

Our work on an initial "beta version" of the LEBOK Guide was initiated in November 2024, driven by a recognized lack of clarity regarding the broad scope of learning engineering as a process and practice. Learning Engineering Toolkit and the work of the IEEE ICICLE provided a foundational corpus of knowledge and definitions.

The development was inspired by established professional domains, specifically software engineering and project management, which utilize authoritative "Body of Knowledge" guides to frame their respective professions. LEBOK adopts a similar structure as the Software Engineering Body of Knowledge (SWEBOK), utilizing Knowledge Areas (KAs) to categorize areas of professional expertise, and further subdividing the knowledge areas into topics and subtopics.

An initial set of knowledge areas were developed using an iterative process. An AI model was trained with content from key publications recognized by the community, including the Learning Engineering Toolkit: Evidence-based Practices from the Learning Sciences, Instructional Design, and Beyond (Goodell & Kolodner, 2023) and a variety of the author's co-authored papers. The list of knowledge areas were then compared to category lists previously developed by the IEEE ICICLE CCC SIG. The author edited the list of knowledge areas and developed a set of topics and then reviewed the combined lists for gaps. Finally, the previously trained LLM was used to identify potential sub-topic definitions. When available, previously published definitions were used to specify the scope of each subtopic.

Although the initial "beta" edition used some aspect of the learning engineering cycle as defined in the learning engineering process (Kessler et al, 2023), it primarily represented iteration within cycles of the Creation stage of the process. The intent is that the wiki can support additional nested iteration as a "team sport" by the learning engineering community. We expect that the complexity and breadth of the learning engineering knowledge areas can be addressed through modularization (Barr, 2022) and nested cycles of iteration (Craig, 2024). And that the wiki interface can be further iterated upon using human-centered design (Thai, 2023) for enhanced usability.

12 Knowledge Areas (KAs)

The "Beta" edition of the LEBOK Guide is organized into 14 distinct Knowledge Areas, reflecting major themes from the IEEE definition of learning engineering and the Learning Engineering Toolkit.

- KA 1: Learning Engineering Basics: Defines the scope, professional disciplines, and the necessity of cross-domain collaboration.
- KA 2: Human-Centered Design Foundations: Covers empathy-driven research, prototyping, and iterative design.
- KA 3: Learning Sciences Foundations: Explores cognitive, social, and motivational aspects of learning, as well as domain-specific pedagogy.
- KA 4: Engineering Foundations: Focuses on systems thinking, control theory, and empirical methods like A/B testing.
- KA 5: Learning Engineering Models and Methods: Discusses agile development, data-informed design, and design patterns.
- KA 6: The Learning Engineering Process: Outlines the lifecycle from challenge identification to implementation and feedback.
- KA 7: Data Instrumentation: Details the infrastructure for collecting learning event data and interoperability standards.
- KA 8: Learning Analytics: Covers predictive modeling, AI, machine learning, and ethical considerations.
- KA 9: Lean Agile Methodologies: Applies frameworks like Scrum and Kanban to learning engineering projects.
- KA 10: Operations & Project Management: Addresses budgeting, scaling, and resource management for complex learning projects.
- KA 11: Professional Practice: Outlines certification pathways, key competencies, and ethics.
- KA 12: The Learning Engineering Enterprise: Focuses on shifting organizational structures from traditional L&D to learning engineering.

Technical Implementation and Standards

To ensure the Guide is functional within the modern digital ecosystem, it is released in several formats:

Interactive Wiki:

The wiki platform allows for quick reference of a single KA, Topic, or Subtopic. The platform hosted at lebok.wiki also will support continuous community review and iterative development.

Figure. 1.

LEBOK.wiki was soft-released in December 2025.

Machine Readable Format (IEEE 1484.20.3-2023):

The Guide is implemented in a machine-readable format (IEEE Sharable Competency Definition). This will allow the body of knowledge to be linked directly to learning resources, digital credentials, and job descriptions with a globally unique identifier and return machine-readable data objects for use in these systems.

PDF Publication:

The human readable downloadable PDF provides a portable and printable representation of the Guide.

Open Source Software Tools:

The author and software development partner also released the software developed to populate the wiki as open source under an Apache 2.0 license at <https://github.com/jgoodell2/LEBOK>.

Ethical Considerations

The LEBOK Guide places a heavy emphasis on ethical practice (11.2 Ethical Considerations in Learning Engineering), particularly regarding data privacy and the responsible use of data (8.5 Ethical Considerations in Learning Analytics) AI in

learning analytics (8.5.3 Responsible Use of AI in Learning Analytics).

Sustainability and Future Considerations

The initial development and hosting of the LEBOK Guide Wiki was supported by an author-initiated GoFundMe campaign (Goodell, 2025). The funding pre-paid hosting costs through 2026.

Future versions of the Guide are intended to be community-maintained with the goal of a community-vetted release in 2027. The wiki currently supports features for commenting and discussions, however, before turning on those features, to maintain site fidelity, a governance and discussion moderation process will need to be established. Additional funding could support development of a sustainable community governance model, recruiting volunteer knowledge area editors and community moderators, and initial community facilitation towards a community-vetted and improved future release.

The author is exploring models of sustainable funding to keep the resource free and open indefinitely. For example, the site hosting could be sponsored by programs that offer education, training, and certification of learning engineering in exchange for listing ads for their programs on the site. Premium sponsors could be prominently placed on as an ad rotation on the main landing page or on targeted knowledge area, topic, and subtopic pages. Companies that offer tools and services to learning engineering professionals could likewise promote their offerings on the site.

Conclusion

The Guide to the Learning Engineering Body of Knowledge represents a significant milestone in the professionalization of learning engineering. By establishing an initial framework of knowledge areas, topics and subtopics for further development through interdisciplinary collaboration, it provides the learning engineering academic and professional communities with a tool for further defining learning engineering as a professional practice. It may be used to define training and credentialing of learning engineering professionals. In addition to human-readable definitions, the machine-readable resources support direct linking metadata from training offering metadata, learning resources, and digital credentials, including links to skills definitions and rubrics.

References

- Barr, A., Dargue, B., Goodell, J., & Redd, B. (2022). Learning engineering is engineering. In J. Goodell & J. Kolodner (Eds.), *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond* (pp. 125–152). Routledge. <https://doi.org/10.4324/9781003276579-8>
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale* (pp. 115-132). <https://doi.org/10.59668/2109.21735>
- Goodell, J. (2025). Help Publish the Learning Engineering Body of Knowledge. Go fund me page. <https://gofund.me/9af2764f3>.
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning Engineering at a Glance. *Journal of Military Learning*, <https://www.armyupress.army.mil/Journals/Journal-of-Military-Learning/Journal-of-Military-Learning-Archives/Conference-Edition-2023-Journal-of-Military-Learning/Engineering-at-a-Glance/>

Goodell, J., & Kolodner, J. (Eds.). (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Routledge.

Goodell, J. (in press). Learning Engineering Body of Knowledge.

IEEE ICICLE. (n.d.). International Consortium for Innovation and Collaboration in Learning Engineering. ieeicicle.org.

H. Washizaki, eds., Guide to the Software Engineering Body of Knowledge (SWEBOK Guide), Version 4.0, IEEE Computer Society, 2024; www.swebok.org.

Kessler, A., Craig, S. D., Goodell, J., Kurzweil, D., & Greenwald, S. W. (2023). Learning engineering is a process. In J. Goodell & J. Kolodner (Eds.). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Routledge.

Thai, K. P., Craig, S. D., Goodell, J., Lis., J., Schoenherr, J. R., & Kolodner J. (2023). Learning Engineering is Human-Centered. In J Goodell (Ed.), The Learning Engineering Toolkit (pp. 83-124). Routledge.

Totino, L., & Kessler, A. (2024). "Why did we do that?" A Systematic Approach to Tracking Decisions in the Design and Iteration of Learning Experiences. The Journal of Applied Instructional Design, 13(2) <https://doi.org/10.59668/1269.15630>



A Tiered Framework for Educational Event Data Documentation: Synthesizing Principles and Addressing Gaps

Xin Wei, Yenda Prado, Jeremy Roschelle, & Pati Ruiz

data documentation

educational event data

Learning Analytics

learning engineering

Open Science

Digital learning platforms generate millions of fine-grained behavioral events, yet inadequate documentation restricts equitable research use. We systematically reviewed five documentation traditions (survey/DDI, FAIR, AI transparency, learning-analytics standards, psychometrics) and identified gaps unique to educational event data: temporal complexity, nested structures, platform business logic, scale/granularity, and access heterogeneity. Synthesizing cross-cutting strengths, we propose five principles—Transparency, Accessibility, Usability, Responsibility, Maintainability—operationalized in a 16-item 3-tiered framework designed to enable platforms to self-assess and improve incrementally. Empirical validation through platform piloting and researcher feedback represents essential next steps.

Introduction

Digital learning platforms (e.g., ASSISTments, Khan Academy, i-Ready) log millions of student interactions daily, capturing full problem-solving processes: hint requests, errors, time allocation, and help-seeking. Such fine-grained behavioral data can transform how we understand learning and personalize interventions (Baker & Siemens, 2014). Learning engineering leverages learning sciences, cognitive science, computer science, and "the rapid developments in the data available and the learning analytics methods available to analyze it" (Baker et al., 2022, p. 4) to build educational systems that support high-quality learning (Baker et al., 2022). Realizing this promise, however, depends on adequate data documentation.

Yet many educational datasets remain underutilized due to documentation inadequacies. Current practices, inherited from survey research traditions, fail to address temporal event streams, nested data structures, and platform-specific business logic. This creates inequitable access to research opportunities where well-connected researchers navigate data effectively while early-career scholars and under-resourced institutions face substantial learning barriers. This documentation gap thus compounds existing inequities in education research and threatens reproducibility (D'Ignazio & Klein, 2020; Makel & Plucker, 2014).

This paper uses a learning engineering framework to address three questions: (1) What documentation gaps exist when applying established traditions to educational event data? (2) What principles address data documentation gaps? (3) How can these principles be operationalized into a practical, tiered framework?

Methods

We used a nested learning engineering cycle (Craig et al., 2025; Totino & Kessler, 2024) to address the challenge of making educational data more accessible. Our work follows the iterative Challenge-Creation-Implementation-Investigation cycle (Goodell & Kessler, 2023):

1. Challenge: Understanding documentation inadequacies through systematic review
2. Creation: Synthesizing principles and developing the tiered framework
3. Implementation: Developing platform self-assessment and improvement approaches (future work)
4. Investigation: Evaluating framework impact on accessibility and quality (future work)

This paper documents our work in the Challenge and Creation phases. Our approach is conceptual and synthesis-based, drawing on the research team's collective expertise working with educational platforms and event data. Empirical validation through platform piloting and researcher feedback will be essential next steps.

Our methodology consisted of three phases: documentation tradition review, gap identification, and principle synthesis.

Phase 1: Review of Five Documentation Traditions

We identified five documentation traditions representing diverse paradigms relevant to educational platforms (social science, computer science, measurement science, open science infrastructure). Selection criteria prioritized traditions with: (1) established standards, (2) research data documentation focus, and (3) different epistemological perspectives. For each tradition, we identified canonical standards and widely cited guidance documents (e.g., specifications, technical notes, flagship papers) through targeted searches of standards organizations' repositories (DDI Alliance, FAIR Data Principles, IMS Global, etc.) and reference chaining from prior reviews. We extracted documentation requirements and assumptions into a structured comparison matrix organized by tradition and documentation focus.

Survey Research and DDI (Data Documentation Initiative): Survey research has developed sophisticated metadata standards for documenting questionnaires, variables, and sampling procedures (Vardigan et al., 2008). DDI provides structured XML schemas for capturing variable-level metadata, including question text, response options, skip logic, and universe statements (DDI Alliance, 2020). This tradition excels at documenting static, rectangular datasets where each row represents a respondent and columns represent variables measured at a single time point (Groves et al., 2009).

FAIR (Findable, Accessible, Interoperable, Reusable) Principles: Emerging from open science, FAIR principles provide high-level guidance for making research data discoverable and reusable (Wilkinson et al., 2016). FAIR emphasizes consistent identifiers, machine-readable metadata, standardized vocabularies, and clear licensing (Wilkinson et al., 2016). These principles tend to prioritize general infrastructure and cross-repository interoperability over domain-specific documentation needs (Jacobsen et al., 2020).

AI Transparency Frameworks: Model cards (Mitchell et al., 2019) and datasheets for datasets (Gebru et al., 2018) document machine learning training data and model behavior. These frameworks address algorithmic transparency, training data provenance, performance metrics across demographic groups, and intended use cases (Raji et al., 2020). They emphasize ethical considerations, bias documentation, and limitations statements (Arnold et al., 2019).

Learning Analytics Standards: Technical specifications like xAPI (Experience API) and IMS Caliper define standardized event formats for capturing learning activities (ADL, 2013; IMS Global, 2015). These standards provide controlled vocabularies for common educational events (e.g., "viewed," "completed," "scored") and specify JSON schemas for event structure. They enable cross-platform event aggregation but provide limited guidance on platform-specific business logic documentation (Siemens & Baker, 2012).

Psychometrics Documentation: Measurement research emphasizes construct validity, multiple forms of reliability evidence, and appropriate score interpretation (AERA et al., 2014). This tradition documents measurement instruments, scoring

algorithms, validation studies, and known limitations (Kane, 2013). It focuses on what data means substantively rather than technical access mechanisms (Messick, 1995).

This comparative analysis revealed that while each tradition contributes valuable documentation practices, none adequately addresses the unique characteristics of educational event data, as documented in Phase 2.

Phase 2: Gap Identification for Event Data

Two members of the research team independently applied each tradition's documentation guidance to educational event data characteristics (temporal sequences, nested structures, platform-generated processes), identifying areas where guidance proved insufficient or silent. Through iterative discussion and cross-tradition thematic coding (Braun & Clarke, 2006), we synthesized recurring breakdowns into five critical gaps:

Temporal complexity gap: DDI was developed for documenting survey-based datasets and their instruments, where time is typically represented as discrete collection periods (Vardigan et al., 2008) rather than continuous event streams. Educational platforms generate continuous event streams where temporal ordering, duration, session boundaries, and time-on-task patterns carry substantive meaning for understanding learning. While learning analytics standards (xAPI, Caliper) specify individual event formats, they provide limited guidance on documenting temporal dependencies, sequence patterns, or longitudinal trajectories.

Nested structure gap: Survey documentation assumes flat, rectangular datasets (rows = respondents, columns = variables). Event data are hierarchically nested (e.g., events within attempts within problems within sessions within students within classrooms) with many-to-many relationships. No existing tradition adequately addresses how to document these complex relational structures or specify aggregation rules across levels.

Platform logic gap: AI frameworks document training algorithms but not production business logic generating behavioral data. Psychometric documentation describes instruments but not software systems mediating collection. Educational platforms embed complex rules (e.g., hint sequencing, mastery thresholds, adaptive pathways) that researchers must understand to interpret observed events, yet no tradition provides guidance for documenting these systems.

Scale and granularity gap: Survey research documents hundreds of curated variables; platforms generate millions of events. Researchers need guidance for event subset selection matched to research questions, not exhaustive documentation of every event type. Existing traditions lack frameworks for tiered documentation at multiple aggregation levels (e.g., raw events, session summaries, student trajectories) or decision trees guiding researchers to appropriate granularity.

Access heterogeneity gap: Survey data is released as single files with codebooks. Event data exists at multiple levels accessible through different mechanisms: bulk downloads, APIs, derived datasets, or restricted-access portals. FAIR principles address repository-level access but not the tradeoffs between different data representations, computational requirements for event-level versus aggregate analysis, or how access method shapes analytic possibilities.

Phase 3: Principle Synthesis

We synthesized five documentation principles by extracting underlying goals from the five traditions and adapting them to address the identified gaps. To operationalize these principles into the tiered framework, we: (1) mapped each principle to observable documentation artifacts (e.g., variable definitions, API documentation, example code), (2) ordered artifacts from minimal to comprehensive, and (3) defined three capability thresholds (Baseline, Enhanced, Advanced). The team refined item definitions and tier boundaries, ultimately producing a 16-item framework. External validation with platform developers and researchers remains essential future work.

The five principles are:

Transparency (what data exists and what it means):

Documents not just individual event types but also temporal relationships, business logic generating events, and how raw events relate to derived measures. Synthesizes variable-level metadata (DDI), provenance (FAIR), training data descriptions (AI frameworks), event semantics (learning analytics), and construct definitions (psychometrics)—adapted to address temporal complexity, nested structures, and platform logic gaps.

Accessibility (how data can be obtained and used):

Addresses multiple access pathways (raw events vs. aggregated datasets), technical prerequisites (API authentication, query languages), and computational requirements for large-scale temporal data. Builds on persistent identifiers and machine-readable formats (FAIR), codebook distribution (DDI), API specifications (learning analytics), and dataset availability statements (AI frameworks)—adapted for access heterogeneity.

Usability (ease of understanding and working with data):

Provides onboarding materials for navigating temporal complexity, example analyses for common research questions, and guidance matching events to questions. Draws from questionnaire documentation (survey research), interoperability standards (FAIR), datasheets with use cases (AI frameworks), reference implementations (learning analytics), and scoring guidance (psychometrics)—addressing scale, granularity, and nested structure challenges.

Responsibility (ethical considerations and limitations):

Addresses temporal privacy risks (behavioral patterns may identify individuals despite demographic anonymization), algorithmic fairness concerns (platform logic influences observed behavior), and limitations of observational vs. experimental inference. Integrates consent documentation (survey research), licensing requirements (FAIR), bias and limitation statements (AI frameworks), privacy compliance (learning analytics), and appropriate use guidelines (psychometrics).

Maintainability (keeping documentation current as platforms evolve):

Addresses rapid platform evolution where new features continuously change event generation. Synthesizes version control practices (survey research), metadata versioning (FAIR), dataset update cycles (AI frameworks), schema versioning (learning analytics), and measurement revision tracking (psychometrics).

We operationalized each principle into assessable criteria across three tiers (Baseline, Enhanced, Advanced), recognizing that platforms have varying resources and maturity while researchers have differing documentation needs. This operationalization involved: (1) identifying documentation artifacts or practices demonstrating each principle, (2) arranging these by increasing comprehensiveness, and (3) defining three capability thresholds.

Results

The resulting 16-item framework (Table 1) emerged by decomposing each of the five principles into 2-4 assessable questions; for example, Transparency encompasses variable clarity (Item 1), provenance documentation (Item 2), and limitations disclosure (Item 3). Each item is assessed at three capability levels:

1. Baseline: Minimum documentation enabling basic understanding and appropriate use. Provides essential information but may require substantial researcher effort.

2. Enhanced: Additional documentation supporting efficient research use through structured metadata, example analyses, and clearer guidance. Reduces time-to-insight for researchers.
3. Advanced: Comprehensive documentation supporting reproducibility, equity (lowering barriers for early-career researchers), and sophisticated secondary analyses. Represents documentation best practices.

This structure enables platforms to self-assess documentation maturity and pursue incremental improvement. For example, a platform might achieve the Enhanced tier for most Accessibility questions while remaining at Baseline for Transparency question 2 (e.g., lacking processing pipeline documentation). This granular assessment helps researchers evaluate whether documentation meets their research needs.

Table 1.
Tiered Documentation Framework Rubric

Dimensions	Checklist Questions	Baseline	Enhanced	Advanced
Transparency	1. Are all variables clearly named and described?	Variable names listed with brief labels.	Full definitions with units and value meanings.	Complete semantic mapping or ontology crosswalks with usage examples.
	2. Is the dataset provenance documented?	Data collection dates or versions stated.	Sources and processing steps described.	Full lineage (raw → processed → analytic) with version-controlled scripts.
	3. Are assumptions and limitations disclosed?	Generic caveats noted.	Specific limitations and known data issues described.	Quantified uncertainty or bias linked to validation results.
Accessibility	4. Is there a human-readable documentation package?	README / PDF with summary info.	Structured manual with cross-references and examples.	Interactive or searchable web documentation portal.
	5. Is structured (machine-readable) metadata provided?	None.	CSV or JSON variable dictionary with types and domains.	Metadata validated against a standard (DDI, Caliper, xAPI) and registered via API.
	6. Are example analyses or domain-appropriate demonstrations provided?	None.	One or more static examples appropriate to dataset's main use (e.g., joining tables, simulation scripts).	Full reproducible notebook or API workflow with narrative interpretation.
	7. Is the dataset publicly findable and citable?	No DOI / repository record.	DOI or repository record exists but limited indexing.	Indexed in major catalogs (e.g., Google Dataset Search, Dataverse) with persistent identifiers.

Usability	8. Are data tables organized and relationships clear?	Basic list of tables.	Join keys / relationships explained textually or diagrammed.	Machine-readable schema or entity-relationship diagram with validation.
	9. Are there guides for common research tasks?	None.	Text instructions for typical joins or aggregations.	Step-by-step workflow or template scripts covering end-to-end analysis.
	10. Are data quality checks or diagnostics available?	None.	Missingness / range summaries provided.	Automated validation scripts or dashboards released with data.
Responsibility	11. Are privacy protections or ethical safeguards documented?	De-identification mentioned.	Detailed anonymization / aggregation procedures.	Formal privacy-impact or bias audit (for sensitive data); for open data, clear attribution & ethical use guidance.
	12. Are sampling or bias considerations disclosed?	None.	Narrative description of representativeness limits.	Quantitative subgroup or equity analysis with interpretation.
	13. Is appropriate-use guidance provided?	Generic license text.	Clear do's / don'ts for intended uses.	Ethical-use policy with community or IRB oversight.
Maintainability	14. Is versioning and change history available?	Single version or date label.	Changelog / release notes maintained.	Semantic versioning with archived prior releases and linked documentation.
	15. Is there a support or feedback channel?	Contact email.	Form or issue log for user feedback.	Public issue tracker with transparent resolution workflow.
	16. Is there a plan or cadence for future updates?	None stated (snapshots acceptable if clearly archived).	Informal commitment to periodic review or metadata refresh.	Formal maintenance schedule or automated continuous integration for documentation.

Discussion

This framework advances educational data documentation in five ways. First, it provides systematic synthesis of documentation traditions explicitly adapted for temporal, nested, platform-generated event data. Second, it offers actionable guidance through 16 checklist questions enabling platforms to self-assess maturity, identify improvement priorities, and progress incrementally without complete re-engineering. Third, it promotes equity by setting transparent expectations that help

researchers evaluate whether documentation meets their needs before substantial time investment, reducing advantages currently enjoyed by well-networked researchers. Fourth, it supplies shared stakeholder language. Funders can specify tier requirements in data management plans. Repositories can provide structured feedback using the rubric. Finally, it provides essential R&D infrastructure for learning engineering by reducing time spent deciphering undocumented data, enabling cross-platform comparisons, and establishing shared standards that accelerate the iterative improvement cycles central to learning engineering (Baker et al., 2022).

This framework represents a conceptual synthesis requiring empirical validation. We have not tested whether platforms can reliably self-assess using these items, whether tier distinctions meaningfully differentiate documentation quality, or whether improved documentation reduces researcher barriers. The framework should be viewed as a theoretically-grounded starting point, not a validated instrument ready for high-stakes deployment.

Following learning engineering practice (Baker et al., 2022; Goodell et al., 2023), the next Implementation and Investigation phases involve: (1) piloting with platforms to assess reliability, (2) gathering researcher feedback on which elements reduce time-to-first-analysis, (3) examining relationships between documentation quality and research productivity, and (4) testing generalizability beyond formal learning platforms. These investigations will inform framework refinement, automated assessment tools, and documentation templates. The framework focuses on content and standards; complementary technical infrastructure (e.g., scalable APIs, privacy-preserving access) and policy challenges (e.g., consent models) represent parallel work.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences (IES), U.S. Department of Education, through SEERNet Grant R305N210034 and National Science Foundation (NSF), through Grant 2153481. The opinions expressed are those of the authors and do not represent the views of the IES, the U.S. Department of Education, or NSF.

References

- ADL (Advanced Distributed Learning). (2013). Experience API (xAPI) specification, Version 1.0.0. <https://github.com/adlnet/xAPI-Spec>
- AERA, APA, & NCME. (2014). Standards for educational and psychological testing. American Educational Research Association.
- Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5), 6:1-6:13. <https://doi.org/10.1147/JRD.2019.2942288>
- Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 253-272). Cambridge University Press.
- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*, 3(1), 1-23. <https://doi.org/10.1037/tmb0000058>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training

Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). <https://doi.org/10.59668/2109.21735>

D'Ignazio, C., & Klein, L. F. (2020). Data feminism. MIT Press.

DDI Alliance. (2020). DDI-Codebook 2.5. <https://ddialliance.org/Specification/DDI-Codebook/2.5/>

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92. <https://doi.org/10.1145/3287560.3287561>

Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. *Journal of Military Learning*, 7(1), 46-59.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Wiley.

IMS Global Learning Consortium. (2015). IMS Caliper Analytics® specification, Version 1.1. <https://www.imsglobal.org/spec/caliper/v1p1>

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., ... & Goble, C. (2020). FAIR principles: Interpretations and implementation considerations. *Data Intelligence*, 2(1-2), 10-29. https://doi.org/10.1162/dint_a_00024

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. <https://doi.org/10.1111/jedm.12000>

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, 43(6), 304-316. <https://doi.org/10.3102/0013189X14545513>

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. <https://doi.org/10.1037/0003-066X.50.9.741>

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229. <https://doi.org/10.1145/3287560.3287577>

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33-44. <https://doi.org/10.1145/3351095.3372873>

Siemens, G., & Baker, R. S. (2012). Learning analytics and educational data mining: Towards communication and collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 252-254. <https://doi.org/10.1145/2330601.2330661>

Totino, L., & Kessler, A. (2024). "Why did we do that?" A systematic approach to tracking decisions in the design and iteration of learning experiences. *The Journal of Applied Instructional Design*, 13(2), 277-281. <https://doi.org/10.59668/1269.15630>

Vardigan, M., Heus, P., & Thomas, W. (2008). Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation*, 3(1), 107-113. <https://doi.org/10.2218/ijdc.v3i1.45>

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1-9.
<https://doi.org/10.1038/sdata.2016.18>



IEEE Standards for Scaled Learning Engineering

Jim Goodell

Co-Design

Design Patterns

learning engineering

Models

Modular Open Systems Architecture

Standards

Learning Engineering Toolkit introduces systems engineering principles for scaled implementation of learning solutions. An important mindset for scalable implementation of learning engineering is recognizing learning as a complex non-linear system that includes learners and learning context as part of the system. Engineers use systems models of various degrees of fidelity and design patterns—repeatable solutions to commonly occurring engineering problems. They build resilient, scalable systems using modular components with standard interfaces, defining the operating conditions, constraints, and tolerances for each module and the system as a whole. The Total Learning Architecture is a collection of IEEE standards for learning technology. These standards provide the necessary technical specifications and constraints that allow the modules of learning solutions developed by learning engineering teams to become physically and functionally interchangeable for scaled implementation. This poster will examine how IEEE standards support translation of research and application of learning sciences into scalable learning solutions.

Introduction

Systems engineering plays a critical role in learning engineering (Goodell, 2023). This is particularly relevant to scaled solutions. An important mindset for scalable implementation of learning engineering is recognizing learning as a complex non-linear system that includes learners and learning context as part of the system. Engineers use systems models of various degrees of fidelity and design patterns—repeatable solutions to commonly occurring engineering problems. They build resilient, scalable systems using modular components with standard interfaces, defining the operating conditions, constraints, and tolerances for each module and the system as a whole. The Total Learning Architecture (Smith, 2021) is an approach to building scalable data-informed systems and ecosystems for learning that draws on a collection of IEEE standards for learning technology (IEEE 2022, IEEE 2023, IEEE 2024, IEEE 2025, IEEE n.d., IEEE n.d.). These standards provide the necessary technical specifications and constraints that allow the modules of learning solutions developed by learning engineering teams to become physically and functionally interchangeable for scaled implementation. The poster shown in figure 1 examines IEEE standards used to support modern systems, including intelligent tutoring systems, for scaled learning.

Figure. 1.

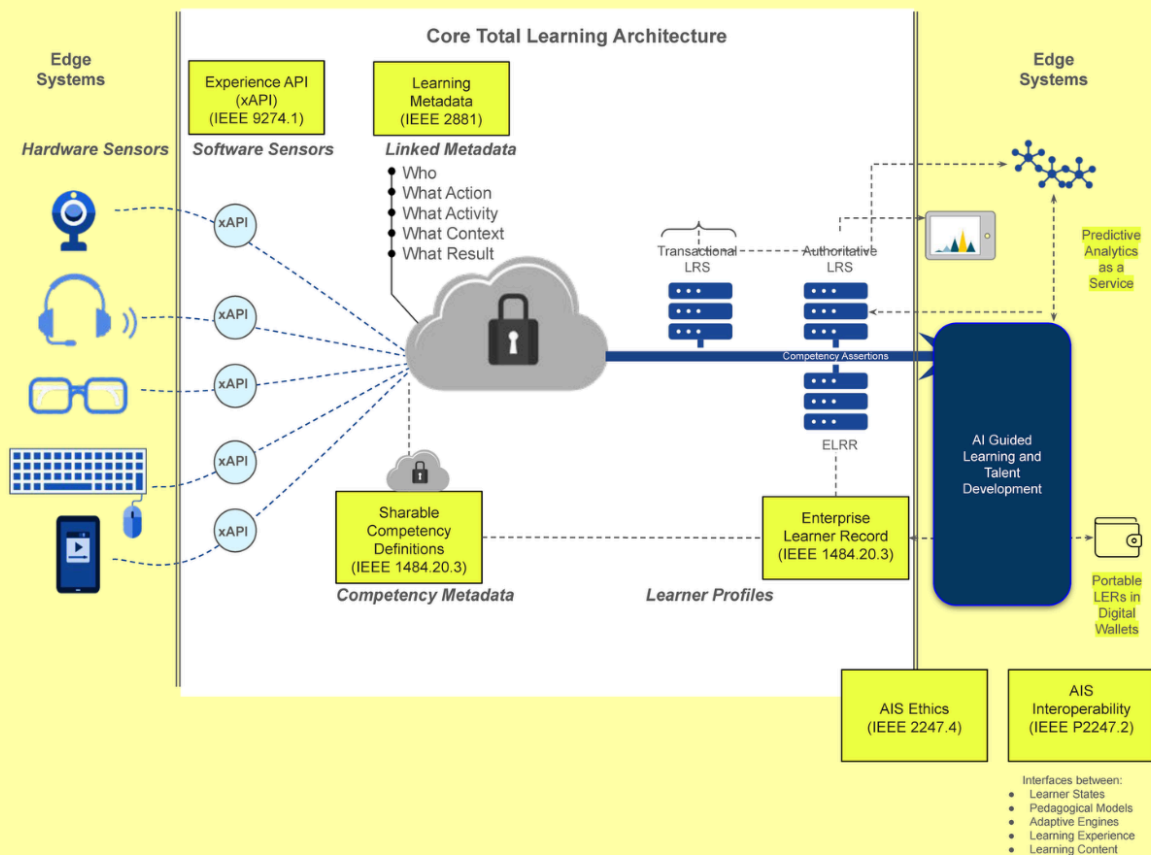
Poster showing IEEE standards in the context of data flows of the Total Learning Architecture.

IEEE Standards for Scaled Learning Engineering

Jim Goodell

INFERable, a Public Benefit Corporation; jim@inferable.app

Learning Engineering Toolkit introduces systems engineering principles for scaled implementation of learning solutions. An important mindset for scalable implementation of learning engineering is recognizing learning as a complex non-linear system that includes learners and learning context as part of the system. Engineers use systems models of various degrees of fidelity and design patterns—repeatable solutions to commonly occurring engineering problems. They build resilient, scalable systems using modular components with standard interfaces, defining the operating conditions, constraints, and tolerances for each module and the system as a whole. The Total Learning Architecture is a collection of IEEE standards for learning technology. These standards provide the necessary technical specifications and constraints that allow the modules of learning solutions developed by learning engineering teams to become physically and functionally interchangeable for scaled implementation. IEEE standards support translation of research and application of learning sciences into scalable learning solutions.



<https://sagroups.ieee.org/ltsc/>

Discussion and Conclusion

The existing standard can be used to support the learning engineering of scalable and cost-effective systems that leverage advanced technologies, such as LLMs. In fact, the introduction of standard structured metadata via MCP in formats defined by standards such as IEEE 2881-2025 and IEEE 1484.20.3-2022 can help overcome current challenges, such as hallucinations and interactions that are less than pedagogically sound when using general purpose LLMs for tutoring. However, more research and standards development is needed to extend the existing standard to fully take advantage of recent advancements in AI, including large language models and agentic AI.

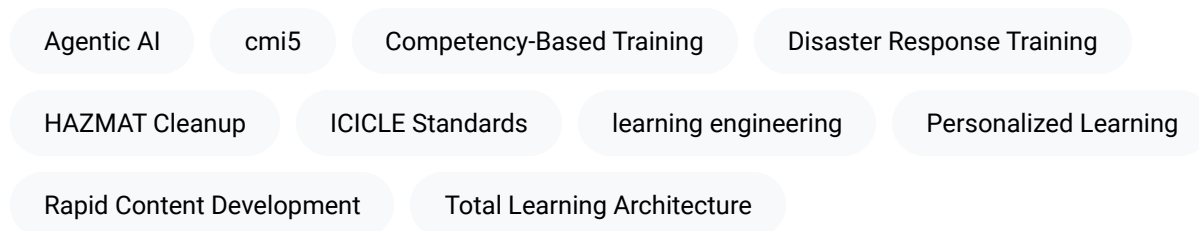
References

- Goodell, J., & Kolodner, J. (Eds.). (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Routledge.
- IEEE. (n.d.). IEEE P2247.2 Draft Standard for Adaptive Instructional Systems Interoperability.
- IEEE. (n.d.). IEEE P2997 Draft Standard for Enterprise Learner Record (ELR).
- IEEE. (2022). IEEE 1484.20.3-2022: IEEE Standard for Learning Technology: Data Model for Sharable Competency Definitions (SCD).
- IEEE. (2024). IEEE 1484.2-2024: IEEE Recommended Practice for Learning and Employment Record (LER) Ecosystems.
- IEEE. (2025). IEEE 2881-2025: IEEE Standard for Learning Metadata Terms (LMT).
- IEEE. (2023). IEEE 9274.1.1-2023: Standard for Experience API (xAPI).
- Smith, B. & Milham, L. (2021). Total Learning Architecture (TLA) Data Pillars and their Applicability to Adaptive Instructional Systems. Report from Advanced Distributed Learning Initiative.
<https://www.adlnet.gov/assets/uploads/TLA%20Data%20Pillars%20and%20their%20Applicability%20to%20Adaptive%20Instructional%20Systems.pdf>



Learning Engineering by Design: An Agentic AI Application for Rapid, Personalized Health & Safety Training in Disaster Response and Hazardous Environments

Henry Ryng & Daniel Suhr



Disaster response and hazardous materials (HAZMAT) cleanup require rapid preparation of diverse, transient workers for site-specific risks. Traditional methods fail to meet tight timelines, creating safety gaps. The Health & Safety Training (HST) Copilot, funded by NIEHS SBIR Phase I (Award 1R43ES037510-01), applies learning engineering principles to embed evidence-based design in an AI platform. Integrating ICICLE best practices and IEEE TLA frameworks, an 11-agent architecture generates personalized primers and assessments from site documents. Development follows a nested learning engineering cycle: human expertise assembly, agentic AI

simulation, evidence/decision tracking (LEED-inspired), and refinement.
Phase I SME rubric evaluation confirmed technical feasibility, instructional quality, and rapid generation, with stakeholder commitment to Phase II.
Future studies will assess learning transfer and safety outcomes.

Introduction

In disaster response (e.g., hurricanes, chemical spills, terrorism incidents) and HAZMAT cleanup, mobilizing workers occurs in hours or days, yet each site presents unique hazards requiring tailored preparation. Workers arrive with varied backgrounds, yet limited contact training time must prioritize hands-on practice over basic hazard review. Inadequate preparation contributes to preventable incidents (OSHA, 2023). Traditional development timelines (6-8 weeks) fail here, while credentialing often relies on attendance certificates offering no insight into competency or currency amid skills decay.

Structured training often fails to address learner variability (varying reading levels, multilingual needs, and prior experience gaps), limiting effectiveness for mobile workforces and emergency responders. Specific contexts include hurricane/wildfire disaster response requiring rapid team mobilization, HAZMAT remediation with site-specific chemical protocols, terrorism/conflict preparedness for emergency response teams, and pandemic response requiring broad multi-language communication of infection control protocols to diverse health worker populations.

The HST Copilot addresses this via learning engineering—a transdisciplinary, iterative process applying learning sciences, human-centered design, systems engineering, and data-informed decisions to create effective, scalable learning solutions (Goodell, Kessler, & Schatz, 2023; Baker, Boser, & Snow, 2022). Unlike traditional instructional design, LE mandates ongoing evidence collection and iteration to improve outcomes systematically.

This project embeds LE "by design": principles become automatic in the platform, benefiting users without requiring their expertise. Development followed a nested LE cycle (Craig et al., 2025): human interdisciplinary teams defined challenges and decisions, agentic AI simulated specialist collaboration, rubric-based evaluation provided evidence, and Learning Engineering Evidence and Decision (LEED) tracking mechanisms documented rationale for traceability and refinement. The platform integrates Industry Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) best practices and guidelines, along with alignment to IEEE Total Learning Architecture (TLA) competency frameworks in the underlying technology, throughout the agent architecture.

Learning Engineering by Design: Human Teams, Evidence Tracking, and Nested Cycles

LE centers on assembling diverse expertise, applying evidence-based practices, and tracking decisions systematically to ensure rigor and iterability (Goodell & Kolodner, 2023; Thai et al., 2023). HST Copilot development mirrored this: a human LE team—including construction supervisors, disaster response coordinators, safety managers, instructional technologists versed in xAPI/omi5, and operational stakeholders—collaborated to define requirements and validate outputs.

Design decisions were systematically documented using Kessler's Learning Engineering Evidence and Decision (LEED) approach (Totino & Kessler, 2024), tracking what evidence informed which decisions and why, enabling retrospective analysis

("Why did we do that?") and continuous refinement. This human process informed the core innovation: an agentic AI "team" simulating specialist collaboration.

Agent 1 (Instructional Designer) defines objectives via backward design. Agent 2 (Training Specialist) classifies knowledge, skills, and abilities (KSAs) using frameworks from MIL-HDBK-29612-2A (U.S. Department of Defense, 2001) and Bloom's Taxonomy (Bloom, 1956)—standards HST stakeholders were already familiar with from their domain expertise. Subsequent agents personalize readability, motivation, visuals, and structure per ICICLE principles (cognitive load management, dual coding, learner engagement). This mimics human LE teams but scales rapidly.

The nested cycle—human expertise → AI simulation → evaluation → refinement—aligns with LE's iterative, data-informed nature (Goodell et al., 2023).

The Training Readiness Challenge in NIEHS Contexts

NIEHS-focused domains (disaster response, HAZMAT remediation, emergency preparedness) demand site-specific, time-critical preparation amid diverse workforces. Challenges include:

1. Timelines: Hours to days for deployment vs. weeks or months for traditional training development
2. Unique hazards: Each incident site presents distinct dangers requiring customized content
3. Learner variability: Reading levels (6th-16th grade equivalents), multilingual needs, prior experience gaps, and frequent inclusion of unskilled workers and day laborers rapidly recruited into post-disaster operations
4. Limited contact time: Must prioritize hands-on practice over didactic review of basic hazards
5. Credential trust: Traditional attendance certificates provide no verifiable evidence of competency or skills currency
6. Skills decay: Mobile and intermittently deployed workers experience competency deterioration without systematic refresher mechanisms

The HST Copilot Solution

Users upload available documents (training materials, site assessments, safety protocols, SOPs); the 11-agent pipeline generates personalized primers (Agents 1-10) aligned to ICICLE standards, then assessments (Agent 11) structured for IEEE Total Learning Architecture (TLA) competency frameworks. cmi5 instrumentation enables analytics for decay detection and competency tracking across deployments. TLA structures produce verifiable digital badges with evidence of specific capabilities, trusted by regulators and employers.

Multi-Agent Architecture with Embedded Standards

Each agent embeds specific learning engineering principles from ICICLE standards, creating a collaborative architecture where outputs build progressively toward complete, standards-aligned training materials:

Agent	LE/ICICLE/TLA Function
1	Extract objectives (backward design)
2	Classify KSAs (MIL-HDBK-29612-2 + Bloom's; TLA alignment)
3	Personalize readability (cognitive load management)
4	Motivational framing (learner engagement)

5	Visual icons (dual coding theory)
6-9	Scenario-based practice items (authentic assessment)
10	Compile with TLA/cmi5 structure (competency evidence)
11	Generate TLA-aligned assessments (performance-based evaluation)

SME review ensures accuracy and safety compliance; human-centered iteration builds stakeholder trust and operational validity.

Phase I Evaluation: Rubric-Based Platform Assessment

Per NIH SBIR Phase I constraints (no human subjects research without IRB approval), evaluation focused on platform technical capability to generate accurate, appropriate pre-training content. A structured 27-criterion rubric assessed primer/quiz quality (content accuracy, clarity, standards alignment, readability appropriateness), platform usability, and overall operational readiness across domains including HAZMAT remediation and disaster response.

Stakeholders (director-level representatives from disaster response and safety organizations) provided high ratings on problem-solution fit (M=4.7/5.0), customization capability (M=4.7/5.0), and overall value proposition (M=4.6/5.0); all (4/4, 100%) committed to Phase II field testing. Preliminary rubric-based evaluation yielded high ratings for content quality (M=4.6/5.0) with zero critical safety errors identified by expert reviewers. Technical validation demonstrated 95% reduction in development time (4.5 hours vs. 6-8 weeks for traditional instructional design approaches), enabling rapid response timelines required for dynamic disaster contexts.

Rubric feedback validated agentic outputs' instructional integrity while identifying areas for enhancement (e.g., assessment distractor realism via future retrieval-augmented generation). This LE-aligned evaluation—systematic evidence collection, decision tracking, and iterative refinement—demonstrates platform feasibility for rapid, personalized training generation.

Addressing Measurement, Trust, and Future Outcomes

ICICLE/TLA frameworks link performance data to competencies via structured evidence (cmi5 instrumentation, scenario authenticity, knowledge checks), supporting construct validity and credibility for regulators and employers. Trust builds through SME oversight, adherence to established standards (MIL-HDBK, Bloom's, ICICLE, TLA), and transparent documentation of design decisions via LEED tracking.

Current work establishes promising infrastructure for rapid, personalized preparation. Phase I validated platform feasibility and content quality through SME evaluation. Phase II will measure actual learner outcomes: transfer to hands-on performance, behavior change in operational contexts, and safety impact (incident/near-miss rates, supervisor performance ratings, competency retention patterns) via IRB-approved protocols in authentic disaster response and HAZMAT environments.

Conclusion

HST Copilot demonstrates LE excellence made automatic: nested cycles combining human expertise with agentic AI simulation, systematic evidence and decision tracking via LEED, and embedded learning sciences standards (ICICLE, TLA) democratize rigorous instructional design for underserved high-risk contexts. By embedding learning engineering excellence into automated processes, the platform democratizes access to rigorous instructional design for underserved populations—

disaster workers, mobile construction crews, and emergency health responders—who traditionally lack such resources. By advancing rapid, verifiable preparation, it supports safer disaster response, HAZMAT operations, and emergency preparedness in contexts where training failures have life-or-death consequences.

Acknowledgments

This work was supported by the NIEHS SBIR Phase I (Award 1R43ES037510-01). The authors thank beta stakeholders for evaluation participation and the ICICLE and TLA communities for standards development that enabled this work.

References

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*. <https://doi.org/10.1037/tmb0000058>
- Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Longmans, Green.
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale* (pp. 115-132). <https://doi.org/10.59668/2109.21735>
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning Engineering at a Glance. *Journal of Military Learning*, 7, 42-52.
- Goodell, J., & Kolodner, J. (Eds.). (2023). *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond*. Taylor & Francis.
- Occupational Safety and Health Administration. (2023). *Commonly used statistics*. U.S. Department of Labor. <https://www.osha.gov/data/commonstats>
- Thai, K. P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. (2023). Learning Engineering is Human-Centered. In J. Goodell (Ed.), *The Learning Engineering Toolkit* (pp. 83-124). Routledge.
- Totino, L., & Kessler, A. (2024). "Why did we do that?" A Systematic Approach to Tracking Decisions in the Design and Iteration of Learning Experiences. *The Journal of Applied Instructional Design*, 13(2). <https://doi.org/10.59668/1269.15630>
- U.S. Department of Defense. (2001). MIL-HDBK-29612-2A: *Instructional Systems Development/Systems Approach to Training and Education, Part 2: Acquisition*. Department of Defense.



Scaling Interleaved Mathematics Practice

Matlen, B. J.

Interleaving

learning engineering

Mathematics Education

Interleaved practice involves mixing different problem types during study rather than blocking them by category. It has been identified as a robust learning strategy that enhances both discrimination learning and long-term retention. Despite compelling laboratory evidence and successful classroom efficacy trials demonstrating large learning gains, interleaved practice is relatively uncommon in mainstream mathematics curricula. This disconnect between research and practice exemplifies a persistent challenge: learning principles validated in controlled settings often fail to make their way into classrooms. The present contribution reports on preliminary results of a cluster randomized study testing the efficacy of interleaved mathematics in a large, U.S. based sample. It also briefly discusses the inherent complexity in translating learning principles to authentic contexts.

Introduction

One product of cognitive science research are principles of learning - teaching and study strategies that have been shown to enhance learning in a range of settings and ages in both classroom and lab tests. One such principle is interleaved practice,

which involves mixing problems of different types during study rather than blocking them by category, and distributing problems of the same type over time. Interleaved practice has been found to improve both discrimination learning and long-term retention and a recent cluster randomized trial indicated substantial learning gains for interleaved practice over blocked practice in middle school mathematics (Rohrer et al. 2020). Still, interleaved practice remains relatively uncommon in mainstream mathematics curricula (Rohrer et al. 2020).

In the present study (ongoing), we are testing the efficacy of interleaved practice relative to blocked practice in a large, cluster randomized study. The study is designed to both replicate and extend the study conducted by Rohrer and colleagues (2020). Specifically, like Rohrer et al, the present study tests the efficacy of interleaving in middle school mathematics classrooms, randomly assigning classrooms within teachers to either an interleaved or blocked control group. However, unlike Rohrer et al, the study expands the sample beyond a single district and state to multiple districts and states in a U.S. nationally representative sample. The study also expands the intervention from a half year to a full year and implements interleaved practice on a digital platform - ASSISTments - which increases the intervention's ability to be implemented at scale in authentic contexts.

Learning principles like interleaving are typically framed as straightforward strategies, but the implementation of such strategies can be challenging as they require many contextual decisions that are not directly informed by past research (e.g., how many items are optimal within an interleaved assignment? Can students work in pairs? etc.). To instantiate the intervention in the present work we engaged in a structured design process guided by a logic model. The logic model is a structured description of the intervention informed by existing research, theory, and practical constraints of the use context, that can be revised and iterated upon in formative tests. Table 1 displays this working model.

Table 1. Logic Model of Interleaved Practice in the present study

The logic model of interleaved practice includes intervention components informed by prior research, materials and implementation features that define fidelity, and their connection to the interleaving intervention that this study is designed to replicate.

Intervention Component	Study Materials	Instantiation in Practice	Relation to past efficacy study
Interleaving problems of different types within assignments to support discrimination learning	Assignments created in ASSISTments that each consist of grade level appropriate problems that are interleaved	Students complete assignments that contain topics they've had prior instruction.	Core feature, held constant
		Students attempt each problem independently, writing out the corresponding solution steps	Core feature, held constant
		Students complete problems in ASSISTments	Design feature, varied
		Students complete assignments in class	Design feature, held constant

Revision of incorrect solutions after each assignment to support comprehension	Visual tool for showing problems and corresponding solutions, such as a whiteboard or screen	Teachers discuss problems with solutions available visually	Core feature, held constant
		Students instructed to correct both their work and answers	Core feature, held constant
		Teachers collect and verify students corrected their errors	Design feature, varied
Spaced practice of problems of same type to support long-term retention	Ordered assignments that consist of problems of the same type spaced across the assignments	Assignments are completed in the order of the study procedure, which are distributed over time so that long-term retention benefits from the spacing effect	Core feature, held constant

One value of this logic model-based approach for replication is that it makes explicit the relationship between implementation features that are considered “core” to the intervention and therefore held constant between this and the prior study, and design features that are not theoretically critical and that can therefore differ from past work to best align with the context. Early formative testing identified several design features that were varied to maximize feasibility and fidelity for this sample and context. This logic model-based approach to learning design shares many similarities to other learning engineering approaches such as the Nested Learning Engineering cycle (e.g., Craig et al, 2025) that involve understanding the problem in context, developing, testing, and revising the working solution.

Method and Results

This work reports on preliminary results of the ongoing study. The present sample (representing two of three planned cohorts) includes 813 sixth and seventh grade students across 51 classes, 29 teachers, 15 schools, and 8 districts. Random assignment occurred at the classroom level, with classes blocked within teachers. The post-assessment, developed collaboratively with the original efficacy study author, consisted of 12 isomorphic items to those in the review assignments, administered four weeks after the final intervention to measure robust learning. Hierarchical linear models accounting for students nested within classes revealed that students receiving interleaved practice ($M = 0.09$, $SD = 1.03$) outperformed blocked practice students ($M = -0.08$, $SD = 0.96$) on the delayed post-assessment, but this difference was just shy of statistical significance at $p < .05$, $F(1, 41.81) = 3.75$, $p = .05$, Hedges' $g = 0.18$. This work shows promise for the potential of interleaving to be effectively applied in authentic education contexts at a broad scale.

Acknowledgments

This research is funded by the U.S. Dept of Education through grant R305R220012 to WestEd and does not reflect the views of the Dept. I thank the many WestEd colleagues, University collaborators, education professionals and students who have supported this work.

References

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C. N. (2020). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology*, 112(1), 40.

Rohrer, D., Dedrick, R. F., & Hartwig, M. K. (2020). The scarcity of interleaved practice in mathematics textbooks. *Educational Psychology Review*, 32(3), 873-883.

Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale* (pp. 115-132). <https://doi.org/10.59668/2109.21735>



Teacher Education through Learning Engineering: An Action Research on Faculty Transformation

Kursat Cagiltay & Funda Alptekin

Action Research

Case Study

Digital Transformation

Faculty of Education

learning engineering

Teacher Education

This study aims to examine the digital transformation process conducting in the faculty of education through an action research approach. The process began in the beginning of 2025 with comprehensive needs analysis and continued with focus group interviews with key stakeholders. Based on the findings, the learning engineering approach has been adopted to the curriculum of faculty of education as a guiding framework. Accordingly, professional development seminars have been conducted for faculty members, courses have been revising, and the design and implementation of four new courses have been planned for the initial phase. The study follows a qualitative research design utilizing action research within a single case study framework. As the transformation of the faculty continues on the basis of the learning engineering framework, this study aims to provide information about the process and the outcomes so far.

Introduction

In today's world, technological changes and societal expectations are rapidly transforming the scope of the teaching profession and the approach to teacher education. The education system and the teaching profession were shaped to meet the needs of the First Industrial Revolution and designed to suit the economic and social structure of that era. A standardized, hierarchical, and teacher-centered structure, similar to a factory layout, was adopted for schools. This system aimed to cultivate a disciplined workforce with basic skills for a mass-production-based economy. However, in the age of digitalization, artificial intelligence, automation, and globalization brought by the Fourth Industrial Revolution, this educational model can no longer meet the requirements of the modern world (OECD, 2025a, UNESCO, 2021).

Today, a world order based on the knowledge economy necessitates raising individuals who prioritize skills such as problem-solving, creativity, collaboration, and critical thinking (World Economic Forum, 2025). Especially with the rise of artificial intelligence, the requirements for 21st century are becoming prominent, allowing each student to learn at their own pace and according to their needs. However, existing education systems, operating on the industrial paradigm of the 19th century, are proving insufficient in equipping students with the competencies required by this era. The teaching profession and the education system must be updated with new paradigms such as individualized learning, digital literacy, and continuous learning. For this, utilization of the learning engineering approach may be a solution to make educational processes more dynamic, personalized, and effective.

According to UNESCO's Global Report on Teachers (UNESCO & International Task Force on Teachers for Education 2030, 2024) teaching profession has challenges in terms of status and sustainability of profession. The report highlights a critical global shortage of teachers, estimating a need for 44 million additional primary and secondary teachers by 2030. A central

cause of these shortages is the declining social status and attractiveness of the teaching profession worldwide. In many countries, the teaching profession is increasingly viewed as less prestigious and less appealing for younger generations. Unattractive salaries, challenging working conditions, and increasing workloads discourage candidates from both entering and remaining in the profession.

Whereas as it is stated in the report (UNESCO & International Task Force on Teachers for Education 2030, 2024), teachers are fundamental to ensuring high-quality education. Therefore, all teachers should receive adequate, relevant and pedagogically sound training aligned with system curricula and contextual needs. However, many teacher preparation systems currently fall short. Many new teachers report feeling unprepared for diverse classrooms, particularly those involving mixed abilities, multilingual students, or vulnerable learners. Moreover, the lack of adequate training in ICT competencies is another challenge. Looking forward, UNESCO emphasizes the need for a new social contract for education that elevates teaching as a collaborative, autonomous, research-informed profession. Teachers are positioning as creative guides who use technology to enhance learning and communication.

Similar to other countries, same problems in teaching profession has been arising in Turkey (OECD, 2025b). The teaching profession is showing a tendency to lose its attractiveness, a situation that threatens the future of education faculties. To train teachers suitable for the requirements of the 21st century, education faculties must transform not only their content and methods but also the fundamental paradigms of the teaching profession. The rapid advancement of technology has necessitated a rethinking of the structure of faculties of education and the processes of teacher education. Traditional teaching roles need to be replaced by technology-integrated, data-driven, and learner-centered roles. In this context, digital transformation refers not only to technological infrastructure but also to pedagogical and structural changes. This study aims to present how a digital transformation process was designed and is been implementing within TED University Faculty of Education.

At this point, learning engineering offers a compelling and systemic pathway to address the multifaceted challenges facing the teaching profession. Learning Engineering is defined as a “process and practice that applies the learning sciences using human-centered engineering design methodologies and data-informed decision making to support learners and their development” (Goodell & Kolodner, p10, 2023) By integrating learning sciences, human-centered design, data analytics, and technology, learning engineering provides structured processes that strengthen both teacher preparation and school-level instructional ecosystems. This approach equips pre-service and in-service teachers not only with pedagogical knowledge but also with future-oriented competencies such as problem-solving, evidence generation, learning analytics, and innovative technology use. “Teaching methods, subject areas, technologies, and data-driven algorithms will change the character of their jobs. Their roles will evolve from “sage on the stage” (a more passive approach focused on information transfer) to “guide on the side” (an active mentor who helps scaffold knowledge and offer coaching across mental, social, emotional, and physical domains)” (Goodell & Kolodner, p405, 2023). Moreover, learning engineering’s iterative design cycles (analyze–design–prototype–test–iterate) foster teacher autonomy, collaboration, and continuous improvement—features that directly align with UNESCO’s vision of a lifelong learning ecosystem for educators. In doing so, learning engineering can enhance the quality of pre-service teaching education programmes, improve professional development, reduce the stress associated with technology integration, and strengthen teachers’ sense of efficacy and motivation. Ultimately, the approach contributes to rebuilding the attractiveness, status, and sustainability of the teaching profession.

Method

This study adopts an action research approach within a single case study design (Fraenkel, Wallen & Hyun, 2012). The researchers took an active role in the planning, implementation, and evaluation of the process. This paper describes the first nine months of ongoing study. The case is TED University’s Faculty of Education. The TED University is a private nonprofit institution dedicated to an English-language liberal arts model in Turkey. Faculty of Education offers bachelor’s degrees in Early Childhood Education, Primary Education, Mathematics Education, English Language Education, and Guidance & Psychological

Counseling, alongside graduate and doctoral programs. With robust accreditations, modern pedagogical and psychology labs, a strong emphasis on practice through K-12 Schools, and supportive systems like the Center for Teaching and Learning, the Faculty actively cultivates innovation, student-centered learning, double majors/minors, Erasmus+ exchanges, and scholarly research (Turkish Education Association University, n.d.).

The faculty's transformation process is being examined through action research (Fraenkel, Wallen & Hyun, 2012). In this multifaceted study, the planning and implementation phases are progressing simultaneously. The observation and reflection phases have not yet been reached. In the first nine months, data were collected through focus group interviews, document analysis, and meeting notes. The collected data were analyzed using thematic analysis.

Focus Group Interviews

In the initial phase, focus group interviews were conducted with stakeholders including academics, preservice teachers, and representatives from schools and edtech private sector. The main question for which the answer was looked for is "How can faculties of education be restructured to train teachers who can keep pace with the digital transformation?". Based on the main question four sub questions were asked to respondents in each focus group:

1. How do you foresee the teaching profession changing in response to rapid technological advancements? What will be the role of teachers and students?
2. How do you envision the transformation that Faculties of Education will undergo in order to train teachers who can keep up with this rapid change in technology?
3. What kinds of needs might arise during this process and what solutions can be offered to address these needs?
4. When structuring the technological transformation of Faculties of Education, which national and international policy documents, standards, qualification frameworks, and different practice examples (examples of updated teacher training programs) can be used as a basis? Why?

There were 10 participants in the focus group with academics in Türkiye, 8 in the focus group with academics abroad, 8 in the focus group with private sector representatives, and 15 in the focus group with students. Participants were reached with convenience sampling. The student group consists of preservice teachers from various universities in Ankara, including students from elementary education (five 4th year, two 1st year), elementary mathematics education (one 4th year) early childhood education (five 4th year), guidance and psychological counseling (two 4th year). The academics abroad were graduates of faculty of education in Turkey and were pursuing their studies in the United States, Germany and the United Kingdom. The academics in Turkey were affiliated with different universities, representing educational faculties and center for teaching and learning.

All interviews were recorded, transcribed and thematically analyzed to explore perspectives on the future roles of teachers and students, the structure of faculties of education, and the nature of teaching and learning processes. Table 1 shows the result of the thematic analysis.

Table 1. Results of thematic analysis

Themes		Participants			
		Student	Sector	Academician (Abroad)	Academician (Turkey)
Faculty of Education	Micro-Credential	x	x	x	x
	Academic Training	x	x	x	x

	International Cooperation Internship		x	x	x	x
	Interdisciplinary Work		x		x	x
	More Internships and Practices		x	x		
	Alternative Measurement and Evaluation		x	x	x	
	AI-Powered Application Area				x	x
	AI-Powered System			x		
	Support Centers				x	
	New Courses	Use of Artificial Intelligence in Education	x	x	x	x
		Virtual Reality	x			
		Technology Integration	x			
		Entrepreneurial	x			
		Digital Technologies			x	x
		Data Analysis- Data Education- Data Literacy		x	x	x
		Coding			x	
		Teaching Productivity Tools			x	
		Ethics-Digital Ethics		x	x	x
		Neuroscience for Faculty of Education			x	
	Social Skills					x
The Role of the Teacher	Guidance		x	x	x	x
The Role of the Student	Individualized Learning		x			x
	Social Skills Development		x			
	Managing Your Own Learning Process			x	x	

Education and Training Processes	Learning in the Classroom	x		
	Virtual, Hybrid, Face to face	x	x	x
	Flipped Learning			x
	Simulation, AI		x	

Deciding of the Approach: Learning Engineering

Based on the results and a comprehensive literature review, the learning engineering (Goodle & Kolodner, 2022) approach was adopted as the most suitable framework for the faculty of education's transformation. Looking ahead to the 2025–2026 academic year, our vision is to redefine teacher education, talent development, corporate training and lifelong learning through Learning Engineering so that graduates are fully prepared to leverage evidence-based methodologies, emerging technologies, and design thinking in diverse professional contexts. This initiative aims to prepare the next generation of educators and learning professionals not only as content experts but as learning engineers—equipped with technological, pedagogical, and ethical competencies to meet the demands of modern education and training.

Key Elements of Our Learning Engineering Approach:

- Interdisciplinary Curriculum:

Faculties of education must move beyond traditional pedagogical competencies and equip pre-service teachers with interdisciplinary 21st-century skills such as data literacy, digital pedagogical competence, design thinking, and human-centered design. Modular and flexible course structures should encourage students to take courses from fields such as engineering, information technologies, social sciences, and psychology. STEAM-based learning, design-driven problem solving, and soft-skill development should also be integrated into the curriculum.

- Integration of Advanced Technologies

Teacher education programs should embed advanced technologies—including artificial intelligence (AI), augmented and virtual reality (AR/VR), learning analytics, and coding—within existing courses and expand them through new courses and micro-credential pathways. Eye-tracking for classroom management, material development, and analysis of learner–teacher–technology interaction should be used to support professional growth. Hybrid and flexible learning environments should be strengthened through modular synchronous and asynchronous learning opportunities. All theoretical and practical courses are being enhanced through AI, virtual and augmented reality, smart robotics, and wearable technologies to create more effective, efficient, and engaging learning experiences.

- Expansion into Lifelong Learning and Talent Development

Faculties of education should embrace a lifelong learning perspective by offering flexible learning pathways to K-12 teachers. Micro-learning–based certifications should be created to support continuous professional development in areas such as data literacy, digital pedagogy, ethics, leadership, and creative problem solving. These pathways should nurture long-term talent development across the teaching profession.

- Graduate Programs and Certification Pathways

New graduate programs should align with international competency frameworks and offer specialization opportunities in educational technology, learning engineering, digital ethics, and education for learners with special needs. Certification pathways aligned with the European Qualifications Framework (EQF) should be established. Portfolio-based assessment, process-oriented evaluation, and data-driven assessment practices should shape these programs.

- European Union Projects and International Collaboration

Faculties of education should actively participate in EU-funded initiatives (e.g., Erasmus+, Horizon Europe, MSCA projects) to build international partnerships, support student and faculty mobility, and promote innovation-driven educational practices. Collaborative projects in AR/VR, STEAM, digital ethics, and learning engineering should be encouraged, with systematic mechanisms for knowledge and resource sharing.

- Global Partnerships

Strategic collaborations with leading institutions should be developed to establish dual-degree programs, co-taught courses, international internships, and joint research. These global partnerships will help pre-service teachers cultivate globally competitive teaching competencies and broaden their access to international professional opportunities.

- Institution-Wide Participation

Transforming teacher education requires an institution-wide approach that extends beyond education faculties. Cross-faculty collaborations, joint course designs, integration with university technology centers, and continuous professional development for academic staff should form the foundation of institutional digital transformation and curriculum renewal.

- Micro-Projects for Social Impact

Teacher education programs should integrate community-based micro-projects that allow pre-service teachers to design solutions to real societal problems. Through these projects, future teachers gain experience in creating social impact, responding to authentic needs, and engaging meaningfully with communities.

Faculty Development

As one of the results of the focus group and as noted in the literature (Judge & O'Bannon, 2008), faculty members who will be implementing the new approach need training to familiarize them with and embrace the learning engineering approach and to enhance their technological competencies. Therefore, Professional development seminars have been organized, conducted, and will continue to be held regularly throughout the process. Some of the seminars are listed below:

- The Use of Artificial Intelligence in Education
- Efficiency and Productivity Training with Artificial Intelligence
- Course Design and Evaluation Training with Generative Artificial Intelligence
- Introduction to Learning Engineering
- The Nature of Engineering and Its Place in STEAM Education
- IEEE ICICLE Learning Engineering Working Group Meetings (IEEE Standards Association, 2025)
- The Use of Eye Tracking Technology in Education
- Augmented Reality Tutor (ART.)
- Developing Future Learning Leaders: Corporate Academy Expertise

Enhancing Courses Offered in Terms of Technology Integration

As part of the digital transformation process undertaken in the faculty, faculty members with strong technological competencies have been identified in each department to review the courses within their department and to lead the department-level process of technology-enhanced curriculum transformation. These faculty members form the task force responsible for the revision process. This structure aims to ensure that the process is developed rapidly, effectively, and through coordinated collaboration across the entire faculty. One faculty member from each department has been designated as a "task force" coordinator. Under the leadership of these coordinators the digital transformation process was initiated by selecting two courses from each department. The departments and selected courses are listed below:

- Elementary Mathematics Education
 - Mathematics Teaching
 - Calculus I
- Elementary Education
 - Fundamental Mathematics in Primary School
 - Character and Values Education
- Early Childhood Education
 - Teaching Practice
 - Children and Media
- English Language Teaching
 - Teaching Reading Skills
 - Teaching Speaking Skills
- Educational Sciences
 - Educational Psychology
 - Instructional Principles and Methods
- Guidance and Psychological Counseling
 - Individual Counseling Practice I
 - Psychology of Learning

Following the identification of courses, meetings were held within each department to discuss how these courses could be enriched in collaboration with the task force representatives. These meetings focused on course syllabi, assignments, and tools used in the courses. Department-specific tools and methods were also discussed. A range of technology integration strategies were explored and planned during the meetings, including AI applications, VR-AR technologies, interactive content, digital materials, assessment tools, eye tracking, classroom-based or virtual simulations, and web 2.0 tools. Below is a summary of selected initiatives discussed:

- Proposing an e-portfolio system to allow students to showcase the work they produce
- Use of eye-tracking to gather data during micro-teaching activities
- Revise rubrics to include items that encourage effective integration of technology
- Guidelines for AI use in assignments.
- Specific tools and their applications across disciplines.

- Flipped learning strategies and use of tools like ART (İren, n.d), NotebookLM, H5P
- Integration of augmented and virtual reality technologies
- Use of virtual client simulations in counseling programs or virtual classroom scenarios in primary education courses.

Throughout the following semester, the proposed implementations have been applied in the courses listed above. During this process, data has been collected, observations have been conducted, and the effective and less effective aspects of the implementations will be identified. After the data is gathered and analyzed, the status of technology integration in the courses will be evaluated, and necessary revisions will be made accordingly. The enhancement of other courses in the faculty in terms of technology integration will begin based on the feedback obtained from the initial implementations.

Educational Forum: Learning Engineering

In May 2025, TED University organized an international education forum whose theme was Learning Engineering, bringing together scholars, practitioners, and policymakers from around the world (Türk Eğitim Derneği, 2025). The Forum aimed to provide a scientific foundation for redesigning teaching and learning processes, emphasizing effectiveness, inclusivity, and sustainability. By bringing together diverse stakeholders, the event served as a platform for collective thinking and collaborative innovation. The forum opened a multifaceted discussion on how to structure learning processes in a more effective, inclusive, and sustainable manner. The Forum included discussions on the evolving role of the teaching profession, digital competencies, instructional design knowledge, and technology-supported pedagogical skills. In this context, models of teacher preparation were examined through the lens of learning engineering and how they may be transformed with about 2000 participants. To our knowledge, this event was the largest and most widely attended gathering ever held globally dedicated specifically to Learning Engineering.

Offering New Courses

Considering that learning engineering is an interdisciplinary process and practice, all courses offered in the faculty of education were examined from this perspective. In the Faculty of Education curriculum, several courses related to instructional technologies, the use of AI in education, and mathematical instructional design were already offered, including:

- EMT 370: Data-Driven Mathematics Instructional Design
- ED 320: Exploratory Data Analysis for Learning Environments
- ED 310: Trends and Emerging Practices in Educational Technology
- ED 330: Digital Classroom Design and Management
- ED 340: Advanced Digital Learning Material Development
- EMT 304: Technology in Mathematics Education
- ED 430: Artificial Intelligence in Education
- ED 305: 21st Century Skills for Teachers

In addition to these courses offered in previous semesters; to support the learning engineering approach some additional courses were determined to open for the first following semester. The first decision was to offer an under graduated level course titled Introduction to Learning Engineering. This course was designed as a case-based course focusing on the foundations of learning engineering, principles and stages of human-centered design methodologies, and core concepts from cognitive science and learning sciences. Students taking this course would learn the fundamental pillars of learning engineering and apply them in a practical, project-based case study.

The second course is Learning Experience Design. This practice-oriented course integrates learning engineering with human-centered design principles, combining UX and UI design methodologies. Students would produce videos, interactive learning materials, and visuals, while also using wireframing tools. The course aims to provide foundational yet applied knowledge for creating effective learning experiences.

In alignment with these two courses, the decision was made to open Educational Data Science, a course constituting a key component of learning engineering. Through this course, students would gain comprehensive knowledge of data analysis processes, enabling them to analyze the data they collect during the design and implementation stages.

In addition to these courses, it was decided to introduce an e-portfolio course that all students in the faculty would take. This course will house all products students create during their studies in educational faculty. Every student who begins a degree program in the faculty will take this course and will showcase all their academic outputs throughout their education. The e-portfolio system, integrated with the LMS, will allow students to continue sharing their work even after graduation. In this way, they will have the opportunity to demonstrate their competencies and strengthen their prospects in the job market.

Conclusion and Recommendations

This ongoing study presents an in-depth exploration of the digital transformation process within the Faculty of Education in TED University in the light of learning engineering approach. Although the observation and reflection phases are yet to be completed, the planning and implementation stages have yielded significant progress in the first year. Through focus group interviews, document analysis, and meeting notes, valuable data have been gathered and analyzed using thematic analysis, providing a robust foundation for data-informed decisions.

Key initiatives have included professional development seminars for faculty members, the ongoing enhancement of existing courses with a focus on technology integration, and the planning and introduction of new, future-oriented courses aligned with the learning engineering approach. Furthermore, the organization of a large-scale forum, engaging approximately 2,000 participants, has demonstrated the institution's commitment to inclusive dialogue and collaborative reform.

The early outcomes suggest that a structured and participatory transformation process—guided by learning engineering principles and supported by action research—holds promise for redefining the pedagogical and institutional landscape of teacher education. As the study progresses into the observation and reflection stages, it is expected to yield deeper insights into the effectiveness and sustainability of these initiatives, ultimately offering a model for faculties of education undergoing similar transformation efforts.

Contemporary educational systems face unprecedented challenges driven by technological transformation, shifting learner needs, and increasing demands for personalized learning. In this context, traditional models of teaching—primarily based on knowledge transmission—prove insufficient. Learning engineering emerges as a multidisciplinary field that integrates educational technology, computer science, data science, human–computer interaction (UX), cognitive science, and pedagogical design to optimize learning through scientific and technology-supported methods. This approach offers a transformative framework for redefining the teaching profession.

Learning engineering focuses on designing, implementing, and evaluating learning experiences using systematic and evidence-based processes. Learning engineers:

- Develop instructional materials and digital tools,
- Analyze learning data to evaluate instructional effectiveness,
- Design personalized learning pathways,
- Integrate emerging technologies such as AI, AR/VR, and learning analytics,
- Contribute to the creation of adaptive and inclusive learning environments.

This expanded professional profile positions educators as strategic designers and innovators rather than solely as content deliverers. The integration of learning engineering into teacher preparation reshapes the professional identity of teachers. Teachers evolve into learning architects who design instruction tailored to individual learner needs, moving beyond the

traditional role of information provider. Through learning analytics and AI-supported tools, teachers analyze student performance, identify specific learning challenges, and design personalized learning plans. They integrate AR/VR, artificial intelligence, gamification, and other emerging technologies to enrich learning both inside and outside the classroom. Teachers develop and apply innovative pedagogical tools and methods that optimize learning processes. Teachers support students not only in academic contexts but also in preparing them for continuous learning throughout life. Moreover, they incorporate insights from neuroscience to plan learning environments aligned with students' cognitive and emotional development. Furthermore, teachers guide students in the ethical and effective use of AI, data science tools, IoT, eye-tracking technologies, and AR/VR systems and cultivate learners' digital literacy, critical thinking, and responsible technology use. Based on these new roles and responsibilities of teachers they become moreover an instructional designer, a data-driven decision maker, an educational technology specialist, a problem solver and innovator, a lifelong learning guide, a critical thinking and digital ethics facilitator and a technology mentor.

To conclude, adopting learning engineering principles offers substantial benefits for both teachers and educational systems. By enabling teachers to better understand individual learner needs, learning engineering supports the design of personalized and differentiated learning experiences that place the learner at the center of instruction. At the same time, the integration of data science and technology-supported design processes increases the efficiency and effectiveness of teaching, allowing educators to make more informed decisions and streamline instructional practices. This shift also enhances the professional prestige of the teaching profession. As teachers move from a traditional knowledge-transmission role to that of a "learning engineer," they become recognized as strategic designers and analytical thinkers, which strengthens the societal perception and status of their work.

For education systems, the transition toward learning engineering may offer transformative advantages. It redefines the teaching profession through the use of scientific and technological tools that optimize learning processes, contributing to greater recognition and attractiveness of the field. Teachers who acquire learning engineering competencies also benefit from increasingly diversified career pathways, gaining opportunities to work in educational technology companies, consultancy organizations, international research projects, and innovation-oriented educational initiatives. Moreover, learning engineering enhances teachers' ability to respond effectively to the demands of the digital era and the Fourth Industrial Revolution by supporting more adaptive, technology-rich, and future-focused educational practices.

Ultimately, integrating learning engineering into teacher education represents a paradigm shift that can introduce deep and systemic innovation into educational systems. Teachers are positioned at the center of 21st-century learning as designers, data analysts, technology leaders, and guides of lifelong learning. Realizing this vision requires structural, conceptual, and pedagogical transformation both within teacher education programs and institutional cultures. Such a transition not only elevates the prestige and attractiveness of the teaching profession but also significantly enhances the quality and effectiveness of education as a whole.

References

- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill
- Goodell, J., & Kolodner, J. (Eds.). (2022). *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond*. Routledge. <https://doi.org/10.4324/9781003276579>
- IEEE Standards Association. (2025). IEEE Learning Technology Standards Committee ICICLE. <https://sagroups.ieee.org/icicle/>
- Iren, D. (n.d.). ARTutor [Web application]. Retrieved from <https://artutor.art/>

- Judge, S., & O'Bannon, B. (2008). Faculty integration of technology in teacher preparation: outcomes of a development model. *Technology, Pedagogy and Education*, 17(1), 17–28. <https://doi.org/10.1080/14759390701847435>
- OECD. (2025a). New professionalism and the future of teaching. <https://www.oecd.org/en/about/projects/new-professionalism-and-the-future-of-teaching.html>
- OECD. (2025b). Results from TALIS 2024 - Country notes: Türkiye. https://www.oecd.org/en/publications/results-from-talis-2024-country-notes_e127f9e2-en/turkiye_754c2c1a-en.html
- Türk Eğitim Derneği. (2025, April 9). TED VI. Uluslararası Eğitim Forumu: Öğrenme Mühendisliği. <https://ogrenmemuhendisligi.ted.org.tr/>
- Turkish Education Association University. (n.d.). Faculty of Education. Retrieved from <https://educ.tedu.edu.tr/en>
- UNESCO. (2021). AI and education: Guidance for policy-makers. UNESCO. <https://doi.org/10.54675/PCSP7350>
- United Nations Educational, Scientific and Cultural Organization & International Task Force on Teachers for Education 2030. (2024). Global report on teachers: Addressing teacher shortages and transforming the profession (ISBN 978-92-3-100655-5; DOI <https://doi.org/10.54675/FIGU8035>). UNESCO.
- World Economic Forum. (2025, January). The Future of Jobs Report 2025: Insight Report. Retrieved from <https://www.weforum.org/publications/the-future-of-jobs-report-2025/>



Understanding Instructor Perspectives and Course Challenges in FSE 100: A Learning Engineering Approach to Improving the Course

Whitney Hansberry, Josh Hales, Mia Dufault, Natalie Carlson, & Scotty D. Craig

Course Evaluation

first-year engineering

Introduction to Engineering (FSE 100) serves as a foundational experience for first-year students, acting as the course to introduce core engineering principles, like design thinking, teamwork, and professionalism. Given its central role, it is critical to understand the perspectives of its stakeholders. Using a learning engineering approach, this study seeks to understand how FSE 100 can be improved, focusing on data, iterative design, and stakeholder insights. Conducted interviews gained insight from the instructor perspective highlighting that a limited credit hours format creates challenges of pacing, workload, and student engagement. These interviews will inform a future survey meant to gain student perspective of FSE 100.

Introduction

Students often perceive introductory engineering courses as unnecessary or disconnected from later coursework. Informal feedback from students enrolled in FSE 100 at Arizona State University (ASU) reflects this pattern, with many citing a misalignment between the course workload and its 2-credit value, as well as difficulty understanding how course content connects to future engineering classes. These perceptions pose a challenge that conflicts with ASU's mission, as unclear course structure and perceived irrelevance can undermine student belonging and equitable access to engineering pathways.

Introductory engineering courses play a critical role in supporting students who may not enter higher education with prior exposure to engineering concepts or practices. When these courses are unengaging, poorly aligned with expectations, or constrained by structural limitations, they can negatively impact student motivation and contribute to lower performance, major changes, or attrition (Ohland et al., 2008). As such, understanding how these courses are designed, implemented, and experienced is essential, particularly in compressed, low-credit formats where instructional choices and structural constraints may have outsized effects on student outcomes.

Guided by a Learning Engineering (LE) approach, this study examines instructor perspectives as part of an iterative, human-centered process aimed at improving FSE 100. LE emphasizes understanding challenges in context, collecting and analyzing multiple sources of evidence, and using that evidence to inform later design decisions (Goodell & Kolodner, 2022; Goodell & Kessler, 2023). Within this framework, the present study represents a nested LE cycle situated within a broader, ongoing effort to improve the course. LE processes are often nested or concurrent, with evidence generated in one phase informing decisions and iterations at other levels of the design system (Totino & Kessler, 2024).

This phase of the project focuses on investigating instructor perspectives to clarify course purpose, instructional priorities, and structural constraints that shape the learning experience. Instructor insights serve as a critical source of evidence for understanding how the course is intended to function and how design decisions are enacted in practice. Findings from this phase are intended to inform subsequent data collection efforts, including the development of a student survey assessing engagement, learning experiences, and perceptions of course effectiveness, as well as future iterations of course design.

Informed by this LE framing, the following research questions shape the research process, analysis of results, and later discussion of improvements:

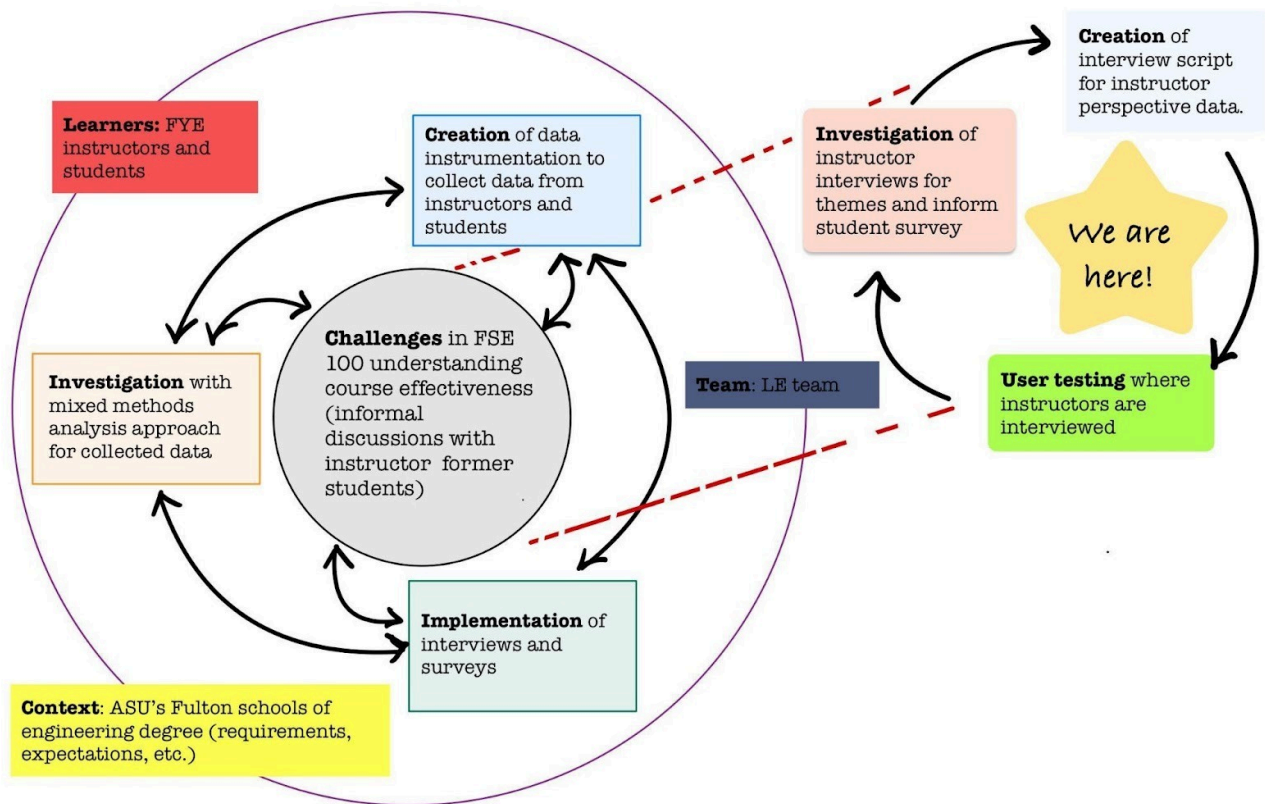
RQ1: Understand instructor perspectives on FSE 100's purpose, learning outcomes, and instructional priorities, and how well the course prepares students for later coursework.

RQ2: Identify how the 2-credit, one-semester format, pacing, workload, and content influence instructors and students.

RQ3: Use findings from faculty interviews regarding course development to design a student survey in a subsequent phase assessing perceptions of course effectiveness.

Figure. 1.

A depiction of the LE cycle showing where our project work fits within the process.



Literature

Informed by a Learning Engineering (LE) approach, the following literature highlights evidence related to first-year engineering course (FYE) design, instructor decision-making, and structural features that influence student experiences and outcomes. A consistent finding in engineering education studies is that early identity and affective development impact student retention in engineering majors. Godwin et al. developed measures of identity in engineering. It found that affective constructs, including interest, identification, and agency, predict the likelihood of students to persist in engineering (Godwin et al., 2016). First-year courses provide the first sustained exposure to engineering, meaning this is when the development of foundational values takes place. Structuring FYE courses to promote positive affective constructs is crucial for building an engineering identity and retaining engineering students.

Instructor perspectives and their instructional choices provide critical evidence for understanding the first-year experience and therefore affect student outcomes. Wahed and Pitterson showed this by investigating the influence of pedagogical knowledge on courses specifically in a FYE context. The research involved conducting interviews with first-year course instructors. It concluded that first-year professors shape student belonging and help-seeking opportunities because of decisions their jobs entail, such as amount of teaming to include, how to intervene in team issues, and how to define “professional” work (Wahed & Pitterson, 2024). This argues that instructor data, including what they notice, value, and put into practice, is valuable evidence to consider when resigning an introduction engineering sequence.

Purpose and structure must also be considered when redesigning a FYE course. Ohland et al. showed that structure is directly linked to retention outcomes. Their study involved multiple institutions, finding that students enrolled in cohesive, structured FYE programs showed higher retention rates than students in programs with fragmented or inconsistent introductory curriculums (Ohland et al., 2008). Therefore, the structure of a course should be considered when designing for increased retention rates. Marra et al. supported this conclusion with its finding that structured first-year experiences impact the

students' sense of belonging. They concluded that women who persevered in engineering attributed their motivation and sense of belonging to courses with supportive expectations and structures early on in their engineering course load (Marra et al., 2012). First-year curricula with a strong, cohesive structure give students greater opportunity to develop important engineering skills, discover relevance, and form connections. This increased belonging creates a framework that encourages persistence from students.

The literature demonstrates that FYE courses are critical in influencing student outcomes. They create a setting where the development of student engineering identity, instructor perspectives, and course structure all interact to support learning, foster belonging, and promote persistence in engineering.

Methods

For the present study, semi-structured interviews lasting approximately 45-60 minutes were conducted by members of the research team. Within the LE framing of this project, instructor interviews serve as an early evidence-gathering step to inform ongoing design decisions and subsequent iterations of the course improvement process. Interviews were conducted over Zoom, with audio-recorded transcripts collected with participant consent. Data from interviews were used to identify key themes related to course effectiveness, instructional and structural challenges, and opportunities for improvement. As this is an ongoing study, findings from this phase will also inform revisions to the interview protocol and the development of a student survey for a subsequent phase.

The interview protocol consisted of five sections (each containing three to five questions) focused on course governance and decision-making, course effectiveness, pacing and workload, student engagement, and course structure.

Participants

Participants were recruited from ASU faculty, including instructors teaching FSE 100 and advisors within the Fulton Schools of Engineering. Recruitment occurred via email. At the time of reporting, three instructors agreed to participate, with one interview completed; recruitment and data collection are ongoing.

Data Collection & Analysis

Each interview involved one participant and two researchers, with one researcher serving as the primary interviewer and the other documenting observational notes and managing recording. Interview transcripts were de-identified and edited for clarity prior to analysis.

Following editing, transcripts were analyzed using a hybrid thematic analysis approach combining inductive and deductive coding, consistent with established qualitative methods for demonstrating rigor (Fereday & Muir-Cochrane, 2006). Inductive coding was conducted by one researcher to identify emergent themes, while deductive coding was conducted by a second researcher using a codebook aligned with the research questions. This approach enabled comparison across analytic perspectives and supported evaluation of how effectively the interview protocol addressed the study's research questions.

Results

As of now, two interviews with FSE 100 instructors have been conducted and recruitment is ongoing. Analysis of the instructor interviews (P01 and P02) revealed three major themes related to course effectiveness, student experience, and the development of FSE 100 over time.

Theme 1: Instructor Views on Purpose, Challenge, and Student Experience

P01 described FSE 100 as an introductory exposure course that gives students a “taste of all departments of engineering” and helps them begin recognizing engineering communication and seeing themselves as engineers. Coding was highlighted as an area where students often experience unfamiliarity and frustration, but it can also spark interest for those new to the material. The course intentionally requires students to confront “uncertainty,” “ambiguity,” and “iteration,” leading them to face (and often dislike) failure before enjoying the final outcome when designs work. As P01 noted, students often “love the outcome, [but] hate the process.” P02 echoed this, stating the most popular activity among students is the final build as they feel this “sense of accomplishment” and “...to have it complete at the end, [the students] are just full of joy.” These insights will inform student survey phase, particularly with respect to engagement, perceived difficulty, and early engineering identity.

Theme 2: Motivation Barriers Linked to the 2-Credit Structure

Student motivation was the most frequently coded factor, with instructors consistently identifying a misalignment between workload and the course’s 2-credit designation as a primary barrier. Students often questioned, “This is just 2 credits, why am I doing so much?”, which reduced their willingness to engage with or persist through challenging tasks. Instructors described strategies such as gamification, enjoyable in-class activities, and instructional flexibility as effective in mitigating these motivation challenges. Several also noted that increasing the course to 3 credits or adjusting the pacing could better support student motivation. As P02 stated, they “do tell [my class] straight up front, [the class] is a 2-credit hour class that’s really the work of a 3-credit hour class.”

Theme 3: Decentralized Curriculum and Limited Oversight Across Sections

Instructors described FSE 100 as a course with high-level learning outcomes that were collaboratively established by faculty teaching FSE 100 and EGR 101/102 courses, particularly given the absence of ABET-associated learning outcomes for the introductory course. While this shared agreement provides general direction, instructors indicated that curriculum implementation below this level varies substantially across sections. At the section level, instructional decisions were described as largely autonomous, resulting in what one instructor characterized as a faculty “free-for-all,” with limited coordination, oversight, or peer review. Reflecting this concern, P02 expressed a desire for “more oversight ... and more peer review” to support greater coherence across sections. As described by instructors, meaningful information flow across the broader engineering curriculum, from FSE 100 (or EGR 101/102) through mid-level coursework and capstone experiences, primarily occurs when faculty teach multiple courses within the pipeline. In these cases, instructors gain direct insight into how early course experiences align with downstream expectations, highlighting the lack of formal structures to support shared understanding and curricular coherence across the program.

Discussion

Findings from the initial instructor interviews align with prior research highlighting the role of first-year engineering (FYE) courses in shaping students’ early affective experiences, including motivation, frustration, and emerging engineering identity. Instructors described FSE 100 as intentionally exposing students to uncertainty, ambiguity, and iteration, requiring them to engage with unfamiliar tools and confront failure as part of the learning process. While students may initially resist this process, instructors noted that successful outcomes are often associated with strong feelings of accomplishment, consistent with research linking early affective experiences to persistence in engineering (Godwin et al., 2016).

Motivation emerged as a central concern, particularly due to the perceived misalignment between course workload and its 2-credit designation. Instructors reported that this mismatch can reduce student engagement and persistence, a finding consistent with prior work demonstrating how structural features of FYE courses influence belonging and retention (Ohland et al., 2008; Marra et al., 2012). Instructional strategies such as gamification, flexible deadlines, and engaging in-class activities

were described as ways to mitigate these challenges, reinforcing evidence that instructor decision-making plays a key role in supporting productive learning environments (Wahed & Pitterson, 2024).

Instructors also identified decentralized curriculum implementation and limited oversight across sections as a broader structural challenge. Although high-level learning outcomes were aligned with downstream courses, variation in section-level execution and limited coordination may contribute to inconsistent student experiences. From a LE perspective, the findings presented here represent early evidence intended to inform subsequent research design aimed at more fully understanding the FSE 100 context, rather than final conclusions about course effectiveness. This evidence, together with additional instructor interviews and future data collection, will further refine understanding of course structures and instructional practices and inform later suggestions for course design improvements.

Conclusion

This study highlights how structural and instructional features of FSE 100 shape early student experiences in engineering. Preliminary instructor insights suggest that the 2-credit format, pacing demands, and decentralized curriculum implementation introduce challenges related to motivation, workload expectations, and instructional coherence.

Situated within a broader LE process, these findings provide instructor-centered evidence to inform subsequent phases of course improvement. Insights from this phase will guide development of a student survey to capture learner perspectives and support evidence-informed decisions about future iterations of FSE 100, with the goal of strengthening engagement, belonging, and persistence in engineering.

References

- Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods*, 5(1), 80-92. <https://doi.org/10.1177/160940690600500107> (Original work published 2006)
- Godwin, A., Potvin, G., Hazari, Z., & Lock, R. (2016). Identity, critical agency, and engineering: An affective model for predicting engineering as a career choice. *Journal of Engineering Education*, 105(2), 312–340. <https://doi.org/10.1002/jee.20118>
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning Engineering at a Glance. *Journal of Military Learning, Conference Edition (iFEST)*. <https://www.armyupress.army.mil/Journals/Journal-of-Military-Learning/Journal-of-Military-Learning-Archives/Conference-Edition-2023-Journal-of-Military-Learning/Engineering-at-a-Glance/>
- Goodell, J., & Kolodner, J. (2022). *Learning Engineering Toolkit: Evidence-Based Practices from the Learning Sciences, Instructional Design, and Beyond* (1st ed.). Routledge. <https://doi.org/10.4324/9781003276579>
- Marra, R. M., Rodgers, K. A., Shen, D., & Bogue, B. (2012). Leaving engineering: A multi-year single institution study. *Journal of Engineering Education*, 101(1), 6–27. <https://doi.org/10.1002/j.2168-9830.2012.tb00039.x>
- Ohland, M. W., Sheppard, S. D., Lichtenstein, G., Eris, Ö., Chachra, D., & Layton, R. A. (2008). Persistence, engagement, and migration in engineering programs. *Journal of Engineering Education*, 97(3), 259–278. https://depts.washington.edu/celtweb/caee/CAEE_Briefs_PDFs/PersistenceEngagementandMigration_Ohland_JEE08.pdf
- Totino, L., & Kessler, A. (2024). “Why did we do that?” A Systematic Approach to Tracking Decisions in the Design and Iteration of Learning Experiences. *The Journal of Applied Instructional Design*. <https://doi.org/10.59668/1269.15630>

Wahed, S., & Pitterson, N. P. (2024, June). A case study on how instructors' pedagogical knowledge influences their classroom practices for first-year engineering courses. Paper presented at the 2024 ASEE Annual Conference & Exposition, Portland, OR. <https://doi.org/10.18260/1-2-46418>



Human Centered Applications within Learning Engineering



A Learning Engineering Approach to Transforming Teacher Practice Through Co-Designing Science Curricular for Multilingual Learners

Yernat Mnuar, Jie Zhang, & Iftekharul Chowdhury

This practitioner-focused paper describes the application of a learning engineering process to co-design a science unit on chemical and physical changes for 6th grade classrooms. In partnership with a group of science teachers and specialist from highly diverse urban schools, researchers facilitated a two-week summer institute, co-creating a unit anchored in a culturally relevant phenomenon. Drawing on the principles of collaborative design as professional development (Severance, 2022; Voogt et al., 2015), the study investigates how the co-design process can be structured and how it impacts teachers' perceived ability to support science learning of diverse learners. Data from co-design artifacts, classroom observations, and focus group teacher interviews revealed that the process fostered significant teacher agency and provided a powerful context for professional learning. The findings offer a replicable model for using learning engineering to create culturally and linguistically relevant science instruction.

Introduction

The effective integration of Science and Engineering Practices (SEPs) into culturally and linguistically diverse middle school classrooms remains a significant challenge (Lee, 2019). Teachers often lack the resources and professional learning experiences paramount to step beyond scripted curricula and design learning experiences that are both rigorous and receptive to students' cultural and linguistic assets (Garcia & Wei, 2014). Situated in a research practice partnership between university researchers and a grades 6-8 science team in a large urban district, this study applies a nested learning engineering methodology (Craig et al., 2024; Kessler & Totino, 2023) to the teacher-researcher collaborative co-design of science curricula. While the overarching project follows a 4-years engineering cycle (Challenge, Investigation, Creation, Implementation), the annual summer co-design institute represents a focused, iterative micro-loop within the Creation phase. This nested cycle comprises its own investigation (needs assessment), creation (unit adaptation), and user testing (classroom implementation), which allows for responsive, year-to-year refinement of the curriculum to better serve all learners including multilingual learners (MLs). This approach aligns with several high-leverage opportunities for learning engineering identified by Baker et al. (2022), specifically those steered toward supporting diversity, equity, domain-based research, and the integration of human and systemic supports in education. This approach, defined as a human-centered and iterative way to design learning environments, is most effective when it focuses on co-designing with teachers (Edelson, 2002).

Unlike mere collaboration, curricular co-design is a structured process where teachers and researchers jointly define problems and develop solutions, creating a context for profound professional development and curriculum innovation (Voogt et al., 2015). This process fosters teacher agency, positioning educators as authors of curriculum rather than bald implementers (Severance, 2022). Furthermore, informed by asset-oriented pedagogies and inquiry-based science learning, the current study

aims to explore how co-design process can be organized and structured to facilitate co-creating expanded science learning opportunities for all learners including MLs.

This study is guided by the following research questions:

1. How can a learning engineering process be applied to co-design a culturally and linguistically relevant science unit for diverse 6th grade classrooms?
2. How do teachers perceive the co-design process and its effect on their ability to support science learning of diverse learners?

Methods

This study involved four 6th grade science teachers and one specialist from three intermediate schools in one large urban district in Southwest Texas. The district serves a highly diverse student population with 61% of multilingual learners. It is part of a larger federally funded project aiming to use a team-based model to support middle school science teachers to adapt existing science curricular that leverage the communicative practice and strengths of diverse learners. The classrooms included students speaking 5-6 different home languages (other than English) with a wide range of English proficiency. This context made the focus on cultural and linguistic relevance critical. The co-design team, including five teachers, one specialist, and three researchers were engaged in a structured, recurring cycle of Investigate, Create, and Implement across one full academic year. This cycle was embedded within the larger project's Creation phase and was operationalized through a two-week summer institute and monthly Saturday professional development sessions to learn and apply the guiding concepts (contextualized science inquiry, multilingual and multimodal resources, and dialogical science discourse) into the curriculum co-design process through co-design cycles: explore, adapt/design, enact, reflect and modify.

Learning engineering procedure:

We utilized an iterative learning engineering cycle consisting of the key phases:

1. Co-design phase (Summer 2025): A two-week summer institute session laid the conceptual foundation for co-design work. Teachers explored the guiding concepts through hands-on activities, model lessons, collaborative work, and reflections. Then each co-design team selected a unit to adapt for fall implementation. Researchers acted as co-design supporters, facilitating teachers through the process of redesigning an existing unit on chemical and physical change. A key negotiation was connecting the unit in the culturally relevant phenomenon of food (e.g., Texas BBQ) and cooking, which teachers identified as a high-interest topic that could connect to their students' lived experiences.
2. Implementation phase and data collection(Fall 2025): The co-designed unit was implemented by all teachers in their diverse classroom settings. To closely monitor lesson implementation and collect student assessment data, researchers visited each teacher and video recorded the entire 5 lesson unit. This observational data, which included video recordings and structured field notes, along with student artifacts, was specifically focused on capturing student engagement and science sense-making practices. Complementing these observations, co-design meeting minutes, and the iterative lesson plans were collected to document the collaborative process. To capture teacher perspectives, we conducted post-unit group debriefing and individual interviews, focusing on their co-design experience, subsequent shifts in instructional practice, and views on student learning outcomes.

Results

The design and implementation of the Physical and Chemical Change unit for 6th-grade students followed a structured Learning Engineering (LE) process, characterized by iterative, data-informed cycles of design, enactment, and refinement. Our

unit's development can be mapped onto two primary phases: Phase 1 - Summer Co-Design (Challenge & Creation) and Phase 2 - Implementation (Implement & Investigate), with embedded micro-cycles that allowed for continuous feedback and adjustment.

Phase 1: Summer Co-Design - Challenge and Creation through Collaborative Backward Design

The co-design process started during intensive summer meetings with full team attendance, embodying the Challenge and Creation phases of the Learning Engineering cycle. The team first identified the core instructional problem: how to teach TEKS 6.6E (evidence of chemical change) in a way that was accessible, engaging, and culturally sustaining for a diverse student population. Informed by LE principles that emphasize contextualized, authentic problems (Baker et al., 2022), the team selected BBQ as the anchoring phenomenon. This choice emerged from discussions about student interests, cultural relevance, and the potential to illustrate both physical and chemical changes in a familiar, real-world context.

Using a backward design approach, the team started with the TEKS standard and mapped out a 5-day learning progression using the 5E instructional model. Each teacher took ownership of one lesson, facilitating distributed leadership while ensuring coherence through shared review sessions. This collaborative creation process allowed for active modeling of activities, multimedia resources, and multilingual supports as a key element of asset-based pedagogy and guiding concepts.

Phase 2: Implementation - Enactment and Embedded Investigation Cycles

During the September implementation, the unit was taught across multiple 6th-grade classrooms. This phase corresponded to the Implementation and Investigation stages of the LE cycle, with nested micro-cycles occurring daily as teachers collected and responded to student data. Each lesson was engineered to generate formative data:

Lesson 1: BBQ discussions and Skittles activity yielded insights into students' prior knowledge and cultural connections.

Lesson 2: Five hands-on stations produced observation notes, student drawings, and exit tickets that captured conceptual understanding and engagement.

Lessons 3–4: CER writing and recipe analysis provided evidence of students' ability to integrate evidence and reasoning.

Lesson 5: Choice board artifacts allowed for differentiated demonstration of mastery.

Teachers engaged in daily debriefs, a form of rapid-cycle investigation that allowed them to adjust pacing, scaffolding, and language supports in response to student needs. The station-based lesson (Day 2) functioned as a nested LE micro-cycle within the broader implementation. Each station was a mini-experiment: students' interactions with materials (steel wool, vinegar and baking soda, milk and lemon) generated immediate feedback, which teachers used to scaffold understanding and prompt deeper inquiry. The unit's emphasis on family recipes, multilingual discussions, and culturally familiar phenomena exemplifies LE's commitment to equity and diversity (Baker et al., 2022). By leveraging students' cultural assets as foundational resources, the unit engineered more inclusive and accessible learning pathways.

Teacher Perceptions of the Co-Design Process and Its Impact on Multilingual Learners

All three teachers highlighted significant increases in student engagement, which they attributed directly to the phenomenon-driven, culturally anchored design of the units. Ms. Herminia connected this engagement to opportunities for personal and cultural connection, explaining that the lessons allowed students to “share their knowledge about what they have personally experienced too.” Ms. Francisca similarly observed high involvement during multimodal activities such as the interactive word wall, where students “went up there and they placed it themselves after we did the reading. It gave them a little chance to read it, connect.” These observations underscore how the co-design process, specifically the intentional integration of translanguaging spaces and asset-based phenomena, which facilitated deeper intellectual and emotional investment from linguistically diverse students.

Teachers consistently described the co-design process as evolving a sense of ownership and adaptive expertise. Ms. Moira’s reflection - “These lessons have kind of become my babies” - epitomizes the shift from implementer to author. This agency manifested in thoughtful preparation, proactive planning, and real-time adjustment. Ms. Francisca, for instance, rated her preparation as 10 because she had “everything very well-prepared prior... in little cups... to maximize our time.” Similarly, Ms. Herminia emphasized the need to strategically purpose time and resources, stating, “If I have only forty-five minutes, realistically, I know I’m probably at best gonna get 30 to thirty-five.” These reflections illustrate how the iterative codesign cycle promoted teacher ownership, pedagogical adaptation, and responsiveness to contextual constraints as a core objective of learning engineering.

Teachers provided preliminary evidence of student learning, particularly in science content and language development. Ms. Herminia noted that chemical changes were “one of my higher performing standards... The average is 74.7.” She also highlighted growth among emergent bilingual students, sharing an example of a student who “got a 31.3... his second test... he had a 69.2. He made a big jump.” She explicitly connected this progress to the dual-language and translanguaging approach used in the co-designed unit, observing, “I can clearly see that in my students because they are brilliant in both languages.” However, teachers also pointed to areas for improvement, particularly in scientific discourse and reasoning. Ms. Francisca, suggested, “I do think that we could have spent more time on a reading aspect so that they could talk more... definitely more of the reading on the explain day. With more student discourse for them to defend.” This feedback underscores the importance of iterative refinement - a hallmark of the nested learning engineering cycle - to better integrate language and science sensemaking practices.

Discussion

This study reveals that a learning engineering approach, when deeply rooted in co-design and nested, iterative cycles, is a transformative pedagogical and relational practice that repositions teachers as authors, students as cultural and intellectual contributors, and science as a site of engagement and belonging. Our work moves beyond theoretical aspiration to demonstrate how equity-driven design can be systematically engineered into the co-design science lesson planning for multilingual learners.

The co-design process we implemented functioned as a powerful form of situated professional learning, one that refined what Ms. Moira described as lessons that became “my babies.” In a field often constrained by scripted curricula and compliance-oriented professional development, our model enabled teachers to engage in the “messy work” of pedagogical design—negotiating relevant phenomena, core science and engineering practices, language and discourse scaffolding for MLs, and integrating student language and cultural resources. It embodies what Baker et al. (2022) identify as “Better Engineer Learning System Implementation in Schools” as a process that honors teacher wisdom as essential to scalable, sustainable innovation.

The choice to anchor the unit in Texas BBQ and cooking was a deliberate engineering decision emerging from teacher insight and collective community knowledge. This phenomenon served as a cultural and cognitive entryway, facilitating students to connect abstract concepts of physical and chemical change to embodied, familial, and community practices. Ms. Herminia’s observation that students were “captivated”, and Ms. Francisca’s note of their engagement with multilingual word walls

underscore how culturally responsive design is also cognitively engaging design. When students see their familiar languages, recipes, and lived experiences validated as scientific resources, science itself becomes a space of identity assertion, belonging, and intellectual power. This aligns directly with learning engineering's call to "Support Diversity and Enhance Equity" by moving beyond one-size-fits-all interventions toward contextually-grounded, asset-based learning experiences.

Reference

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Journal of Learning Analytics*, 9(3), 1–15. <https://doi.org/10.18608/jla.2022.7555>
- Craig, S. D., Avancha, K., Malhotra, P., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2024). Using a nested learning engineering methodology to develop a team dynamic measurement framework for a virtual training environment. *Proceedings of the International Conference on Learning Engineering*, 1–18.
- Edelson, D. C. (2002). Design research: What we learn when we engage in design. *The Journal of the Learning Sciences*, *11*(1), 105-121. https://doi.org/10.1207/S15327809JLS1101_4
- García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan.
- Kessler, A., & Totino, L. (2023, July 24–26). Reality is much more complex: Understanding and exploring nested learning engineering processes [Conference session]. IEEE ICICLE Learning Engineering Conference 2023, Pittsburgh, PA, United States.
- Lee, O. (2019). Science education for English language learners: A framework for research and practice. *Science Education*, *103*(5), 1239–1266. <https://doi.org/10.1002/sce.21515>
- National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press. <https://doi.org/10.17226/13165>
- Penuel, W. R., Roschelle, J., & Shechtman, N. (2007). Designing formative assessment software with teachers: An analysis of the co-design process. *Research and Practice in Technology Enhanced Learning*, *2*(1), 51–74. <https://doi.org/10.1142/S1793206807000300>
- Severance, S. (2022). Organizing for teacher agency in curricular co-design. *Journal of the Learning Sciences*, *31*(1), 27–59. <https://doi.org/10.1080/10508406.2021.1998412>
- Totino, L., & Kessler, A. (2024). "Why did we do that?" A systematic approach to tracking decisions in the design and iteration of learning experiences. *The Journal of Applied Instructional Design*, 13(2). <https://doi.org/10.59668/1269.15630>



A Learning Strategy Analysis for Guiding the Creation of a Team Training Immersive Reality Environment

Robert F. Siegle, Vipin Verma, Randall D. Spain, & Scotty D. Craig

Immersive Virtual Reality

Learning Strategy Analysis

Team Training

Effective training for military medical personnel requires rapid decision-making, coordinated teamwork, and repeated practice in realistic conditions. To support this need, we are developing an immersive virtual reality triage system integrated with the Generalized Intelligent Framework for Tutoring (GIFT) to enable adaptive instruction, monitor team cohesion, and provide real-time performance assessment. Yet the pedagogical foundations guiding short, iterative, team-based VR training remain unclear. This project applies a Learning Strategy Analysis to identify how principles from gamification, educational burst gaming, extended reality, microlearning, and intelligent tutoring can inform the system's design. Through a targeted literature review and analysis of the VR environment, we identified key heuristic categories, including micro-learning, instructional flow, game environment, roles, narrative problem structure, assessment, and learner needs. The resulting review offers an approach for evaluating and improving instructional design in immersive team-training systems, supporting more effective and scalable training for combat medics.

Introduction

Training medical military personnel is critical to saving lives in emergencies such as a CCP, where rapid and effective care can mean the difference between survival and loss. To support this mission, a virtual reality (VR) experience has been created to provide soldiers with additional means of training (Craig et al., 2024). This training uses the Generalized Intelligent Framework for Tutoring (GIFT), an open-source software platform that enables dynamic training. Designed to enhance the training of combat medics, our team-based training application incorporates tailored instruction and real-time feedback to the learner while providing instructors with metrics on learner performance. However, the creation of such learning materials borrows elements from other forms of pedagogy, and lacks its own distinct identification.

Background

The team-based immersive VR system (Craig et al., 2024). we are developing in GIFT includes an emphasis on short, focused, and repetitive learning exercises. Aspects of gamification, educational burst gaming, gamelets, extended reality, and microlearning aspects are all included within this combined VR and GIFT framework. This review aims to identify the best practices for the components for learning within immersive VR and intelligent tutoring, and propose the creation of a checklist to assess this learning system during development. To better guide development of these areas, we conducted a Learning Strategy Analysis (LSA) (Roscoe & Craig, 2022) to examine similar pedagogical frameworks used in gamification, VR, and

shorter lessons that could be applied to guide the development of the team training exercise in GIFT. The goal of LSA is to identify the best practices within learning strategies used to support an educational system and provide a heuristic checklist to guide the system during development (Roscoe & Craig, 2022; Ville et al., 2021).

Short-form Interactive Learning

Short-form interactive learning environments leverage brief, interactive, and often game-inspired activities to support rapid knowledge, skill acquisition, and engagement across digital training platforms (Craig et al., 2022). Gamification introduces game-like mechanics to learning materials that promote motivation and enhance learning, with serious games having long demonstrated effectiveness across domains (Estellés-Arolas & González-Ladrón-De-Guevara, 2012; Kasurinen & Knutas, 2018). Educational Burst Gaming (EBG) and GameLets extend this approach through fast-paced, repetitive tasks designed to build proficiency (Amresh et al., 2024; Baron et al., 2016), incorporating structured role-based interactions similar to thinkLets (Kolfschoten et al., 2005). Extended reality further enhances the impact of these short-form learning modalities by providing scalable, low-risk, and repeatable training opportunities that may be difficult to provide with traditional in-person exercises (Allen et al., 2016; Djukic et al., 2015; Fuentes, 2018; Siegle et al., 2020). Finally, microlearning principles reinforce these structures through targeted units of instruction and feedback (Hug et al., 2006), though challenges remain in sustaining meta-cognitive reflection and higher-order thinking (Jahnke et al., 2019).

Narrative

Emphasizing the need to balance game design and educational value (Baron, 2017), elements of narration, fidelity, and even the means of skill acquisition all impact learning potential. Lester and colleagues (2013) argue that narrative should be woven into gameplay, even suggesting integrating intelligent tutoring systems (ITS) into a game. ITSs have been known to provide the learner with adaptive content (Graesser et al., 2019). An ITS could develop the story throughout each level without interruption, progressing the narration as the learner reaches key points to feed intrinsic motivation (Lester et al., 2013). Narrative could be further improved by having a virtual human present the narration, which has been shown to increase learning outcomes (Schroeder & Craig, 2021; Siegle et al., 2023). This guidance suggests we could use virtual humans to provide instruction and guidance to trainees in our course, issuing the trainees their standing orders and providing a narrative framework and social component that can guide them through the virtual training environment (Siegle & Craig, 2024).

Fidelity

Simulation fidelity refers to the degree to which a training environment replicates aspects of the real-world system or task it is intended to represent. High-fidelity environments allow for a more realistic experience, fostering deeper learning, motivation, and team performance (Yuzhakov, 2023). Aiming to provide the most realistic virtual experience possible. Within team-based immersive VR systems, fidelity extends beyond environmental realism to include the behavior of non-player characters, teammates, and instructional agents. These elements influence trust, credibility, and social presence, which in turn affect engagement and learning outcomes (Schroeder et al., 2023; Siegle & Craig, 2024).

Learning and Skill Acquisition

Gamification and EBG work best when the core gameplay remains the same from level to level, allowing the user to focus on the increased difficulty or new challenge being tested rather than relearning their role or controls with the game (Baron, 2017). Keeping the gameplay repetitive enough to test the skillset multiple times through play, spaced-out repetition is recommended over continuous repetition of the same skill tested (Amresh et al., 2024). The current GIFT experience aims to achieve this, adhering to principles of learner analytics as users engage with the learning experience (Nguyen et al., 2021).

Methods

LSA is a heuristic evaluation method for learning technology that combines a review of the system to identify learning supports/scaffolds built into the system, a literature review to identify best practices, and heuristic evaluation to determine overlap and potential improvements. This process has four steps: 1) Identify the system's learning strategies; 2) Literature review on each strategy for effectiveness; 3) Comparison between best practice and system; and 4) Providing ratings and recommendations. The current paper provides an overview of Steps 1 and 2. For Step 1, the relevant project and GIFT documentation were reviewed, leading to the identification of seven categories of learning strategies. A targeted literature review was conducted for each category to identify best practices for implementation. A heuristic checklist was created based on the review.

Outcome

The outcome of this review is to understand and apply best practices. Using them to create a heuristic checklist to guide the development of the system. After considering the GIFT-based course experience and reviewing the extant scientific literature we identified seven broad heuristic categories and twenty criteria that should be evaluated across them. Using four different ratings for each criterion, which have been adapted from past heuristic checklists regarding learning technologies (Craig & Schroeder, 2023; Li et al., 2021). Summaries of the heuristics are provided below, with a full list of the rubric located [here](#).

Heuristic Categories

Micro-Learning.

The micro-learning heuristic emphasizes brief, focused instructional units that support efficient skill acquisition without overloading learners. Structured, task-oriented content and repeated exposure foster long-term retention through improved motivation (Nikou & Economides, 2018). Effective micro-learning requires lessons that are engaging, practically relevant, and appropriately scoped to the task at hand (Hug et al., 2006). This approach aligns with cognitive learning principles by reducing overload and maintaining attention by making lessons engaging, practical, and bite-sized (Jahnke et al., 2020).

Instructional Flow.

The instructional flow heuristic refers to the organization and sequencing of learning activities within the immersive environment. Strong implementations provide a clear progression that includes contextual introductions, explicit objectives, opportunities for guided practice, and timely feedback to reinforce performance (Jahnke et al., 2020). In team-based VR scenarios instructional structure should support smooth transitions between phases of action without disrupting immersion, providing structured sequencing through careful guidance (Kirschner et al., 2010). Poorly scaffolded designs may present tasks without sufficient framing or closure, leaving learners uncertain about goals or expectations.

Game Environment Design.

The Game Environment Design heuristic addresses how interface elements and system feedback support learning within the immersive experience. Effective environments include well-timed screen interventions, mission completion indicators, and synchronized task stages (Septiani & Rosmansyah, 2021). These features help learners interpret progress without breaking engagement (de Souza e Silva & Delacruz, 2006). Poor design may overwhelm users or fail to provide meaningful feedback.

Roles.

The Roles heuristic evaluates how clearly and meaningfully player and non-player roles are defined within the system (de Souza e Silva & Delacruz, 2006). Strong implementation provides well-specified player responsibilities, consistent non-player scripted roles, and an appropriate number of distinct roles to support teamwork (Bots & van Daalen, 2005). Inadequate implementations may blur role boundaries or underutilize non-player characters, reducing opportunities for collaboration and coordinated decision-making.

Plot and Problem Situations.

The plot and problem situations heuristic focuses on the structure and realism of the narrative problem space. Effective implementation presents focused, coherent problem situations that reinforce training objectives and reflect realistic operational constraints (Marocco et al., 2015). Systems benefit when narrative is well integrated with the learning context, and seamlessly with task demands, while not linking plot to situational task may include underdeveloped scenarios that fail to meaningfully support learning or engagement (Bots & van Daalen, 2005).

Assessment Heuristic.

The Assessment heuristic addresses how learner performance is evaluated and communicated during training. Strong implementations provide real-time feedback aligned with task performance and incorporate structured assessment opportunities (Septiani & Rosmansyah, 2021). Utilizing built in tools to assess the players navigation of the game and measure their skill level (Bots & van Daalen, 2005). Lower-quality implementations may rely on delayed or minimal feedback, limiting learners' ability to adjust strategies during training.

Learner Needs.

The learner needs heuristic considers whether the system accommodates users' accessibility, social, and knowledge requirements. High-quality implementations support diverse learners through accessible interfaces, meaningful social interaction, and alignment with learners' prior knowledge (Jahnke et al., 2020). Examining the individual needs of the learner, providing the necessary scaffolding and feedback to direct their experience (Roscoe & Craig, 2022). In contrast, weaker implementations may neglect accessibility considerations or fail to support team communication, undermining learning.

Conclusion

Team training for medical personnel is critical to improve skill acquisition, speed, and team cooperation. The VR experience in GIFTaims to be both a means of supplemental training and a way to assess complex team dynamics. In order to achieve these goals, it is imperative that the learning materials created be up to the highest standards and best practices. However, training these highly specialized teams means strict adherence to conventional pedagogy approaches may not be the best choice for material creation. The longer scenarios, complex and changing tasks, and serious nature of the scenarios conflict with microlearning, electronic burst gaming, and gamification elements respectively. These conflicts are only partial, meaning that care must be given when applying best practices to the current gift experience. The proposed checklist focuses on broader assessment items. Enabling for the foundation of a new assessment measure for this new wave of potential learning opportunities.

Acknowledgments

The research described herein has been sponsored by the U.S. Army Combat Capabilities Development Command under cooperative agreement W912CG-23-2-000. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

- Allen, C. J., Straker, R. J., Murray, C. R., Hanna, M. M., Meizoso, J. P., Manning, R. J., ... Hannay, W. M. (2016). Recent advances in forward surgical team training at the U.S. army trauma training department. *Military Medicine*, 181(6), 553–559.
- Amresh, A., Verma, V., & Zandieh, M. (2024). An In-Depth Evaluation of Educational Burst Games in Relation to Learner Proficiency. *Multimodal Technologies and Interaction*, 8(10), 88.
- Baron, T. (2017). An architecture for designing content agnostic game mechanics for educational burst games (Doctoral dissertation, Arizona State University).
- Bots, P. W., & van Daalen, C. E. (2005). GameLets: Taking a playful tack in group support. *Group Support and Negotiation*.
- Craig, S. D., & Schroeder, N. L. (2023, September). Improving online learning ecosystems with science of learning best practices. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 67, No. 1, pp. 270-276). Sage CA: Los Angeles, CA: SAGE Publications.
- Craig, S. D., Siegle, R. F., Li, S., Cooper, N. R., Liu, Y., & Roscoe, R. D. (2022, September). An investigation of the PERvasive Learning Systems impact on soldiers' self-efficacy for self-regulation skills. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 66, No. 1, pp. 742-746). Sage CA: Los Angeles, CA: SAGE Publications.
- De Souza e Silva, A., & Delacruz, G. C. (2006). Hybrid reality games reframed: Potential uses in educational contexts. *Games and Culture*, 1(3), 231-251.
- Djukic, M., Adams, J., Fulmer, T., Szyld, D., Lee, S., Oh, S.-Y., & Triola, M. (2015). E-Learning with virtual teammates: A novel approach to interprofessional education. *Journal of Interprofessional Care*, 29(5), 476–482.
- Estellés-Arolas, E., & González-Ladrón-De-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2), 189-200.
- Fuentes, G. (2018). Real readiness: Marine corps moves to investigate live-virtual-constructive training. *Sea Power*, 61(4), 33–35.
- Graesser A., Hu, X., & Ritter, S. (2019). History of distributed learning. In J. J. Walcutt & S. Schatz (Eds.), *Modernizing learning: Building the future learning ecosystem*. Washington, DC: Government Publishing Office.
- Hug, T., Lindner, M., & Bruck, P.A. (2006). Microlearning: Emerging concepts, practices, and technologies after e-learning, In *Proceedings of Microlearning*. Innsbruck: Innsbruck University Press.
- Jahnke, I., Lee, Y. M., Pham, M., He, H., & Austin, L. (2020). Unpacking the inherent design principles of mobile microlearning. *Technology, Knowledge and Learning*, 25(3), 585-619.
- Kasurinen, J., & Knutas, A. (2018). Publication trends in gamification: A systematic mapping study. *Computer Science Review*, 27, 33-44.
- Kirschner, P. A., Sweller, J., Clark, R. E., Kirschner, P. A., & Clark, R. E. (2010). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist. *Based Teaching Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching*,(November 2014), 37-41.

- Kolfschoten, G.L., Briggs, R.O., De Vreede, G.J., Jacobs, P.H.M., and Appelman, J.H. (2005) Conceptual Foundation of the ThinkLet Concept for Collaboration Engineering. In: Group Decision and Negotiation (forthcoming, accepted for publication).
- Lester, J. C., Ha, E. Y., Lee, S. Y., Mott, B. W., Rowe, J. P., & Sabourin, J. L. (2013). Serious games get smart: intelligent game-based learning environments. *AI Magazine*, 34(4), 31-45.
- Li, S., Siegle, R. F., & Craig, S. D. (2021, September). Case report for the learning science evaluation checklist. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 65, No. 1, pp. 561-565). Sage CA: Los Angeles, CA: Sage Publications.
- Marocco, D., Pacella, D., Dell'Aquila, E., & Di Ferdinando, A. (2015, September). Grounding serious game design on scientific findings: The case of ENACT on soft skills training and assessment. In *European Conference on Technology Enhanced Learning* (pp. 441-446). Cham: Springer International Publishing.
- Nikou, S. A., & Economides, A. A. (2018). Mobile-Based micro-Learning and Assessment: Impact on learning performance and motivation of high school students. *Journal of Computer Assisted Learning*, 34(3), 269-278.
- Nguyen, A., Wandabwa, H., Rasco, A., & Le, A. L. (2021, January). A framework for designing learning analytics information systems. In *Proceedings of the 54th Annual Hawaii International Conference on System Sciences, HICSS 2021*. January 4-8 2021. IEEE Computer Society.
- Roscoe, R. D., & Craig, S. D. (2022). A heuristic assessment framework for the design of self-regulated learning technologies. *Journal of Formative Design in Learning*, 6(2), 77-94.
- Schroeder, N. L., Chiou, E. K., Siegle, R. F., & Craig, S. D. (2023). Trusting and learning from virtual humans that correct common misconceptions. *Journal of Educational Computing Research*, 61(4), 790-816.
- Schroeder, N. L., & Craig, S. D. (2021). Learning with virtual humans: Introduction to the special issue. *Journal of Research on Technology in Education*, 53(1), 1–7.
- Septiani, A. P., & Rosmansyah, Y. (2021). Features, frameworks, and benefits of gamified microlearning: A systematic literature review. In *2021 3rd International Conference on Modern Educational Technology* (pp. 130-135).
- Siegle, R. F., & Craig, S. D. (2024). The voice quality of pedagogical agent impacts learning and agent perceptions. *Journal of Computer Assisted Learning*, 40(5), 2278-2291.
- Siegle, R. F., Roscoe, R. D., Schroeder, N. L., & Craig, S. D. (2020). Immersive learning environments at scale: Constraints and opportunities. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 64, No. 1, pp. 1165-1169). Sage CA: Los Angeles, CA: Sage Publications.
- Siegle, R. F., Schroeder, N. L., Lane, H. C., & Craig, S. D. (2023). Twenty-five years of learning with pedagogical agents: History, barriers, and opportunities. *TechTrends*, 67(5), 851-864.
- Yuzhakov, A. A. (2023). Game Technologies and High-Fidelity Patient Simulation in the Field of Psychology and Medicine. *The World of Games: Technologies for Experimenting, Thinking, Learning: XXIII Professional Culture of the Specialist of the Future*, Volume 2, 829, 45.



Advancing Usability of an Immersive Virtual Reality Team Training Environment within a Learning-Engineering Cycle

Parkhi Malhotra, Vipin Verma, Robert F. Siegle, Kamala Avancha, Kevin Gary, Jamie C. Gorman, Randall D. Spain, Benjamin Goldberg, & Scotty D. Craig

Immersive Virtual Reality

learning engineering

Usability Heuristics

A Virtual Reality (VR) training environment for Tactical Combat Casualty Care (TC3) was built around the Team Dynamics Measurement Framework (TDMF) to address the need for adaptive team training inside a Casualty Collection Point (CCP). Guided by a Learning Engineering (LE) process, we embedded this work within nested LE cycles that emphasized iterative refinement. Two rounds of heuristic evaluation were conducted using a combination of VR and domain specific heuristics to assess the system's usability and its support for team training. This paper focuses on findings from the second heuristic evaluation which contribute to the design of scalable, team VR training aligned with LE principles.

Introduction

Today's military environments are increasingly dynamic, requiring teams to coordinate and perform complex tasks under time pressure and uncertainty. To address this challenge, our team developed a virtual reality (VR) training environment focused on Tactical Combat Casualty Care (TC3), for a casualty collection point (CCP) scenario that incorporates the Team Dynamics Measurement Framework (TDMF; Avancha et al., 2024) to develop and assess team adaptability skills (Craig et al., 2024). Virtual and synthetic environments have a long history of supporting effective training while reducing the risks and logistical burdens associated with live exercises (Andrews & Craig, 2015; Shubeck et al., 2016). However, VR based training systems present unique usability challenges such as inconsistent interactions, limited collaboration, and misalignment between physical and virtual worlds, which can hinder learning outcomes if left unaddressed (Sutcliffe & Gault, 2004; Derby et al., 2024).

To guide system development, we adopted a Learning Engineering (LE) approach, using nested LE cycles to integrate user-centred and domain specific feedback. A key part of this involved conducting heuristic evaluations to assess system usability. Heuristic evaluation is an expert driven usability testing method that identifies the mismatch between design and user expectations, using already established heuristics (Nielsen, 1994). Our first, heuristic evaluation, reported in Malhotra et al. (2025), applied a combination of heuristic for virtual environments (VE) proposed by Sutcliffe & Gault (2004) and the Derby Dozen principles introduced by Derby et al., (2024) using Nielsen's severity rating scale (Nielsen, 1992) to evaluate the impact of each identified usability issue, from minor to critical usability. These include issues in natural engagement and sense of presence, interaction with equipment, avatar height inconsistencies, system responsiveness, limited support for team-based interaction and orientation within the VR environment. These findings were fed back into the LE cycle as design requirements. The development team implemented a series of changes, including improved avatars, button interactions, medical equipment, refined navigation, orientation cues and adjustments to visuals where the key priority for the team was enabling multiplayer functionality to support team-based training. To evaluate the effectiveness of these changes, we conducted a second heuristic evaluation focused on identifying remaining usability concerns that could impact trainee performance.

A Learning Engineering Process

This work describes our nested LE cycles within the development phase, focusing on iterative heuristic evaluations to inform development and incorporate TDMF to train team adaptability skills. The VR system and TDMF was developed following the LE process (Goodell & Kolodner, 2023; Kessler et al., 2023), which included four phases: (1) Challenge, where the problem is examined in context; (2) Creation, where solutions are designed to meet user needs; (3) Implementation, where the solution is deployed and data are gathered; and (4) Investigation, where the collected data are analyzed to evaluate the solution's impact. Nested LE cycles (Totino & Kessler, 2024; Craig et al., 2025) were used to (a) characterize the training challenge and end-user constraints through conversations with military partners and Subject Matter Experts (SMEs), (b) design training solutions using hybrid Cognitive Task Analysis (hCTA) and event-flow diagrams centered on patient care at a CCP and (c) specify perturbations that elicit adaptive team behavior while managing the CCP. Such nested cycles of design and evaluation are a common practice in LE, where smaller iterative loops within the larger phases help refine complex solutions (Avancha et al., 2024; Craig et al., 2025). SME reviews iteratively refined the scenario structure, timing, and scenario narrative so that the VR training would both align with doctrine and create opportunities to observe meaningful changes in team dynamics across successive perturbations.

Method

The evaluation followed a structured protocol adapted from the first evaluation cycle (Malhotra et al., 2025) and heuristics relevant to the VR system were consolidated to assess six key dimensions of VE: Before entering the VR environment (Items=4); Navigation in VR (Items=4); Tasks within VR (Items=4); Feedback and Collaboration (Items=2); Post Task and Exit Scenario (Items=2); and Scenario-Specific (Items=1) (see Table 1). This was applied to the updated version of the VR training environment, which incorporated improvements to the avatar displays, interactions, role-based tasks, visual transitions, and

scenario flow. The purpose of the evaluation was to assess whether usability improvements made in response to the first-round findings effectively addressed previous concerns and to identify any new issues before implementing with trainees.

A team of three (n=3) human factors researchers conducted the evaluation, all of whom were familiar with immersive systems and had prior experience with the VR training scenario. Findings indicate that having expertise in usability evaluation, as well as in the specific interface domain being evaluated, leads to significantly better results. (Derby et al., 2024; Nielsen, 1994). The evaluation was conducted using Meta Quest 3 head-mounted displays. Each evaluator rotated through the three predefined trainee roles in the scenario: Prioritizer, Stabilizer, and Medical Supplier representing the structure of TC3 within the CCP setting. The evaluation included Scenario 1 and Scenario 2, with a 10-minute break in between sessions to minimize VR-related fatigue within a standard lab environment. During the sessions, each evaluator independently took notes on usability breakdowns, interface inconsistencies and confusing elements. The evaluators also had an option to think aloud while they were inside the VE and their observations were recorded. Evaluations were concluded individually to allow each participant to record their observations without influence from others. After completing their sessions, the evaluators convened to synthesize findings, discuss key usability issues, and assign consensus-based severity ratings.

Table 1.
Heuristic Checklist for VR Usability Evaluation

Phase	Heuristic	Description
Before Entering VR	1. Unboxing and Setup	Users should be introduced to the UI, features, and interaction methods.
	1. Instructions	Instructions should provide actionable feedback.
	1. Support for Learning	Cue active objects and provide explanations as needed.
	1. Consistent Departures	Mark and apply design compromises consistently.
Navigation in VR	1. Natural Engagement	Interaction should match real-world expectations.
	1. Sense of Presence	The user should feel present in a real world.
	1. Navigation and Orientation Support	Users should be able to locate themselves and reset positions.
	1. Organization and Simplification	UI should focus on immersive elements, not external controls.
Tasks within VR	1. Task Compatibility	Virtual tasks and object behavior should reflect real-world expectations.

	1. Natural Expression of Action	Allow physical actions; avoid restriction by hardware.
	1. Integration of Physical and Virtual Worlds	Enable effective task completion using virtual tools.
	1. Action and Representation Coordination	User actions and avatar behavior should align with <200ms delay.
Feedback & Collaboration	1. Realistic Feedback	Actions should result in expected and immediate responses.
	1. Team Collaboration	Include landmarks to orient users sharing virtual space.
Post-Task & Exit	1. Comfort	Use should not cause discomfort or fatigue.
	1. Clear Entry/Exit	Entry and exit should be intuitive and communicated.
Scenario-Specific	1. Dynamic Adaptation	The system should respond to unexpected changes or disruptions.

Findings

The second heuristic evaluation identified a moderate number of usability issues, few of which persisted from the initial evaluation despite targeted design updates. Many previously critical problems such as interaction breakdowns, lack of avatar alignment, and successful implementation of multiplayer functionality were addressed. These updates resulted in improved presence, more reliable control and experience and support for role based changes.

However, evaluators ran into a few of lower-severity but instructionally relevant issues. E.g. scenario logic remained partially opaque: certain prompts, such as “Ask Location,” were inconsistently triggered. The scripts for dialogues by non playing characters (NPC) needed an update. This disrupted team coordination and created delays in both the scenarios. Other interface issues included overlapping menus, inconsistent exit wound toggle that hindered task navigation. Some object states such as crates or oxygen tanks did not update consistently across users, which limited shared situational awareness (SA).

Additionally, object behavior remained inconsistent in some cases: thermometers and nasal tubes occasionally clipped through patients or failed to respond to expected interactions, and critical equipment like medical bags (see Figure 1) spawned in inaccessible areas. This issue also highlights the limitations of testing inside a lab environment since spatial orientation remains a challenge, particularly when users removed and refitted headsets mid-scenario. These actions often caused avatar misalignment, floating hands. Instructions specifically have to be given to the participants by the experimenters, which should be a part of the scenario.

Figure. 1.**Scene from the VR training environment showing available medical supplies**

Overall, the second evaluation reflected a notable reduction in the severity of usability concerns compared to the first round, but also emphasized the need for continued refinement in instructional support, feedback systems, and collaborative task alignment. These findings reinforce the importance of iterative evaluation during the LE cycle and provide actionable guidance as the project advances into the implementation phase with participants from the Army testing the system.

Discussion

This study represents a nested LE cycle focused on refining a VR-based team training environment in a CCP setting. The initial evaluation identified critical usability issues in terms of Natural Engagement and Sense of Presence, including absence of user's avatar that informed substantial design changes. These included improvements to avatar height, interaction with objects, and multiplayer functionality. The second evaluation, presented in this paper, assessed the effectiveness of those changes and identified remaining usability concerns that, while less severe, could still affect the system's user experience, team training and coordination.

Findings from the second evaluation reaffirm the value of iterative, expert-driven usability evaluations in complex VR training systems within the LE cycle. Although the severity of issues decreased, gaps in design highlight the ongoing challenge of aligning system functionality with training objectives. These insights directly support the Investigation phase of the LE cycle

and will inform targeted refinements in interface behavior, role-based guidance, and scenario progression. The improvements from this cycle are expected to enhance user experience and the fidelity of analytics derived from user interactions.

The next step in the broader LE process is to transition into the Implementation phase. A formative evaluation with Army trainees will be conducted to assess the system's usability, effectiveness, and team training to generate ecologically valid insights on effectiveness, and usability in operational contexts. These findings will inform further refinement, guide the integration of Generalized Intelligent Framework for Tutoring (GIFT) based adaptive feedback, and support the development of a scalable, data-driven VR training solution grounded in LE principles.

Acknowledgments

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research described herein has been sponsored by the U.S. Army Combat Capabilities Development Command under cooperative agreement W912CG-23-2-000. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

- Andrews, D. H., & Craig, S. D. (Eds.). (2015). Readings in training and simulation (Vol. 2): Research articles from 2000 to 2014. Human Factors and Ergonomics Society.
- Avancha, K., Malhotra, P., Gorman, J. C., Verma, V., Spain, R., Goldberg, B., & Craig, S. D. (2024). Development of team dynamics measurement framework for adaptive teams. Proceedings of the 2024 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC). National Training and Simulation Association.
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). <https://doi.org/10.59668/2109.21735>
- Craig, S. D., Gary, K., Gorman, J. C., Verma, V., & LiKamWa, R. (2024). A synthetic training environment for assessing changes in team dynamics with the Generalized Intelligent Framework for Tutoring. In A. M. Sinatra (Ed.), Proceedings of the 12th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym12) (pp. 97–104). U.S. Army Combat Capabilities Development Command – Soldier Center.
- Derby, S. K., Hughes, C. E., & Archer, R. D. (2024). The Derby Dozen: 12 usability heuristics for AR and MR. Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC). National Training and Simulation Association.
- Goodell, J., & Kolodner, J. (Eds.). (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Taylor & Francis.
- Kessler, A., Craig, S. D., Goodell, J., Kurzweil, D., & Greenwald, S. W. (2023). Learning engineering is a process. In J. Goodell & J. Kolodner (Eds.), Learning engineering toolkit (pp. 29–46). Routledge.
- Malhotra, P., Verma, V., Siegle, R. F., Avancha, K., Gorman, J. C., Spain, R. D., Goldberg, B. S., & Craig, S. D. (2025). Evaluating the usability of a VR team training environment utilizing extended reality (XR) heuristics. Proceedings of the Human Factors

and Ergonomics Society Annual Meeting, 69(1), 1729–1733. <https://doi.org/10.1177/10711813251371033>

- Nielsen, J. (1992). Reliability of severity estimates for usability problems found by heuristic evaluation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 129–130). ACM.
- Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'94), 152–158. ACM.
- Shubeck, K. T., Craig, S. D., & Hu, X. (2016). Live-action mass-casualty training and virtual world training: A comparison. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60(1), 2103–2107.
- Sutcliffe, A. G., & Gault, B. (2004). Heuristic evaluation of virtual reality applications. Interacting with Computers, 16(4), 831–849.
- Totino, L., & Kessler, A. (2024). “Why did we do that?” A systematic approach to tracking decisions in the design and iteration of learning experiences. The Journal of Applied Instructional Design, 13(2). <https://doi.org/10.59668/1269.15630>



Case Study: Learning Economy Foundation Competency Graph

Jim Goodell

Competency Graph

Component

learning engineering

The Learning Economy Foundation's LearnCard is a lifelong learning passport that empowers learners to collect, own, and share their skills and experiences. The Foundation asked INFERable's learning engineering team to develop a prototype skills graph that would (1) integrate into LearnCard to visualize learning trajectories and mastery learning progress, (2) use IEEE standards, (3) display real-time predictions about a learner's likelihood of success on the next attempt at a task based on prior attempts, and (4) be released as open source code. Our team applied a sample of learning sciences and data science methods in an iterative learning engineering process to develop a prototype codebase and web component. The resulting code was published as open source by the client. More research is needed for adapting this approach to other platforms and contexts.

Introduction

In this project, our team applied a sample of learning sciences and data science methods in an iterative learning engineering process (Goodell & Kolodner, 2023) to develop a prototype codebase and web component. The solution:

1. draws a skills graph using IEEE 1484.20.3-2022: IEEE Standard for Learning Technology: Data Model for Sharable Competency Definitions (SCD) (IEEE, 2022),
2. a separate web service predicts the learner's success on the next attempt using a Bayesian Knowledge Tracing method and data instrumented from prior attempts at a skills-based performance task and stored in a Learning Record Store using IEEE 9274.1.1-2023: Standard for Experience API (xAPI) (IEEE, 2023), and
3. updates each gauge in the skills graph with the prediction from the web service.

The resulting code was published as open source by the client (Goodell & Meizen, 2025). The poster in figure 1 shows an image of the resulting graph.

Discussion and Conclusion

The technical approach and open source resource was successfully demonstrated as a reusable resource. In addition to meeting the client's requirements, the modular prediction service used and the graph visualization has been repurposed in a Moodle plug-in that supports mastery practice (INFERable, 2025). More research is needed for adapting this approach to other platforms and contexts.

Figure. 1.

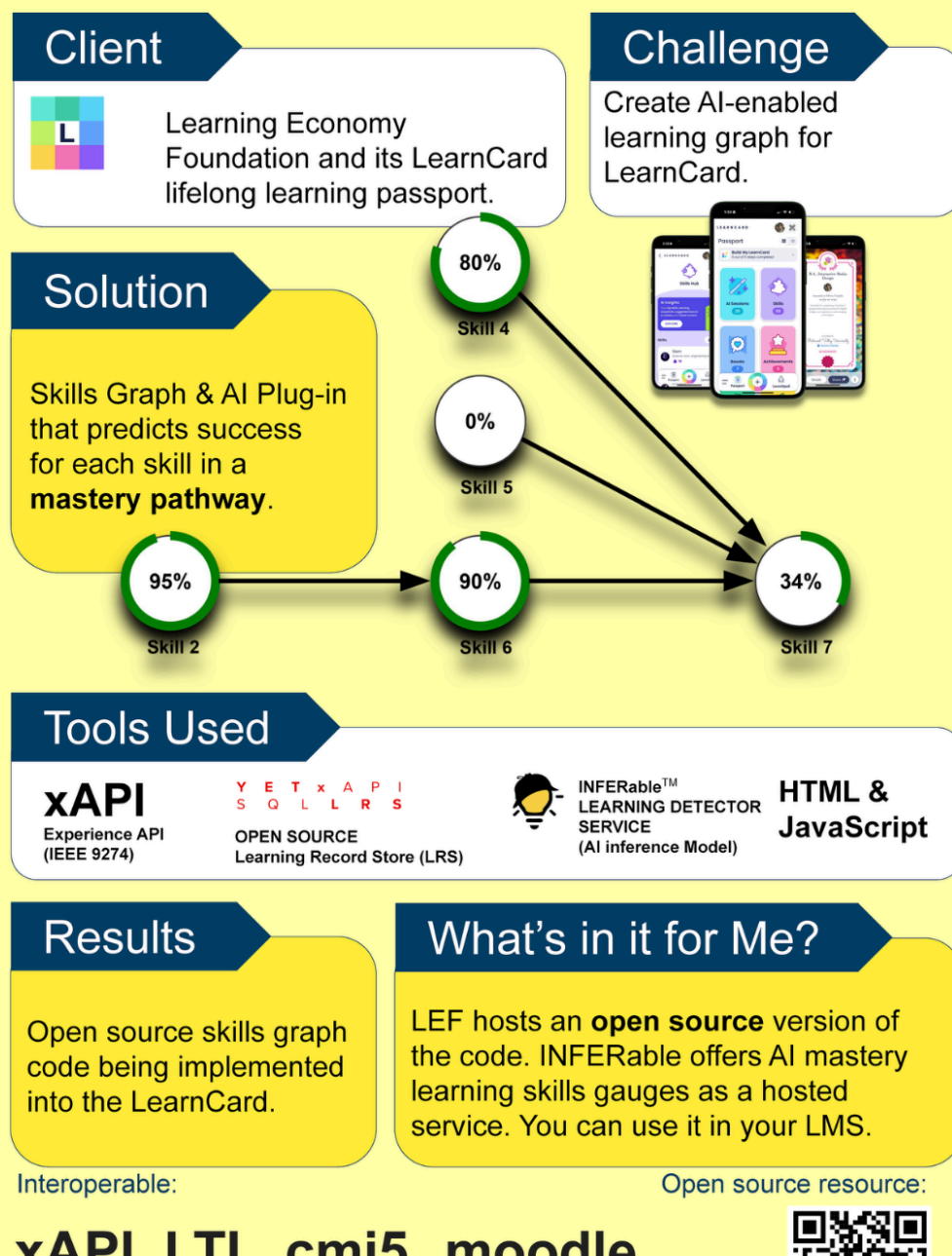
Poster presented at LERN26 showing the Learning Economy Foundation Competency Graph case study.

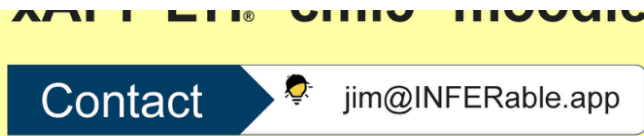
Case Study: Learning Economy Foundation Competency Graph

Jim Goodell

INFERable, a Public Benefit Corporation; jim@inferable.app

In this project, our team applied a sample of learning sciences and data science methods in an iterative learning engineering process to develop a prototype codebase and web component. The resulting code was published as open source by the client. More research is needed for adapting this approach to other platforms and contexts.





inferable.app/use

References

- Goodell, J., & Kolodner, J. (Eds.). (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Routledge.
- Goodell, J., & Meizen, B. (2025). inferable-skills-graph Github repo. Learning Economy Foundation.
- INFERable. (2025). Use on Your Platform. Last accessed 24 January 2026.
- IEEE. (2023). IEEE 1484.20.3-2022: IEEE Standard for Learning Technology: Data Model for Sharable Competency Definitions (SCD).
- IEEE. (2023). IEEE 9274.1.1-2023: Standard for Experience API (xAPI).



Co-Designing a Study Platform for ASU Graduate Students to Enhance Learning Productivity and Performance Through Connection

Fiodesy Gemilang Putri, Valentina Melikhova, Sammy Kwak, Sathkeerthi Subraveti Venkat, & Yuan-Lun Ting

accountability

Collaboration

Learning platform

student motivation

This project explores how ASU graduate students manage focus, productivity, and connection while studying, and how a digital study platform can support these needs. We utilize the challenge and creation phases of the Learning Engineering process to define core learning problems, gather early user insights, and generate design ideas. The findings informed the early concepts for a study connection platform that supports accountability, shared routines, and productive learning behaviors, which we tested through co-design activity. This work sets the foundation for later iterative development and measurement.

Introduction

This report presents our ongoing work in designing a study platform that helps ASU graduate students strengthen their learning productivity through meaningful peer connections. Many students manage demanding schedules and shifting priorities between part-time work and study. As a result, the process of staying focused becomes inconsistent. To address this, we adopted the learning engineering process as our guidance to approach the problem from identification to shaping early design ideas (Totino and Kessler, 2024).

Our work focuses specifically on the challenge and creation phases of the learning engineering cycle. Within both of challenge and creation phases, we conducted nested cycle of learning engineering processes where multiple cycles were needed to understand the challenge before moving forward, and where multiple pieces of the solution need to be created and tested (Craig et al., 2025). In these phases, the goal is to understand the learning problem in context, gather relevant literature, and generate initial solutions that can be refined through future testing. The methods used in these phases help us understand the learning behaviors in the world and build design concepts grounded in evidence.

Objectives and Challenges

Our project goals in the challenge faced were to understand how graduate students manage independent study, what hinders or motivates their focus, and how peer accountability supports learning. These challenges shaped the early design direction for a tool that fosters connections among learning individuals (graduate students at ASU) in an online setting, ultimately enhancing their learning performance.

Methodology

Literature review

The literature highlights the importance of peer connection, motivation, and structure in learning. Research on collaborative study shows that cohesive peer networks and shared regulation processes improve motivation, persistence, and academic outcomes (Berthelon et al., 2019; Haataja et al., 2022). Supportive peer relationships foster accountability, metacognitive interaction, and a sense of belonging, all of which contribute to sustained engagement (Shao et al., 2024).

Motivation research further suggests that both intrinsic and extrinsic factors influence learning behaviors. Gamification elements such as progress tracking, shared goals, and meaningful rewards can increase engagement when aligned with learning outcomes (Schlosser et al., 2023; Smirani & Yamani, 2024; Xu et al., 2021). Studies on group size indicate that while optimal size varies by context, consistency and clear roles often matter more than the number of participants (Connell et al., 2023; Corrége & Michinov, 2021; Zhan et al., 2022).

Learning strategy analysis

To understand how students approach their work, we examined the learning strategies, which include spaced repetition, goal setting, distributed practice, and accountability. Our analysis suggests that a study connection platform should not only facilitate peer interaction but also embed effective learning strategies such as session timers for spaced learning, pre-session planning prompts, and periodic check-ins, so they become a seamless part of the user experience rather than standalone tools.

Prototyping

Based on insights from the literature and early student discussions, we developed low- and high-fidelity prototypes of a study connection platform for ASU graduate students. Prototyping was used as a human-centered design method to refine ideas, communicate concepts, and gather early feedback (Thai et al., 2022). The prototype emphasized core functions such as peer matching, guided study sessions, timers, and progress tracking.

Focus Group Discussion

After developing the prototype, we decided that a focus group discussion would be an effective way to gather qualitative reactions, understand needs and wants, and generate early insights before delving deeper into development (Thai et al., 2022). We conducted a 120-minute in-person focus group with five graduate students using a participatory design approach. Activities included discussions of current study habits, walkthroughs of prototype flows, and a card-sorting co-design exercise. The prototype can be seen on this link: bit.ly/asupeermatch.

Result and Future Research

The focus group session yielded rich insights into graduate students' study behaviors, expectations for study platforms, and experiences with the prototype. Five ASU graduate students participated in a 120-minute moderated session that included: (1) a brief discussion of current study practices, (2) structured walkthroughs and usability tests of three prototype flows, and (3) a card-sorting co-design activity. The findings reveal key barriers to productive study, essential platform features, and design considerations for future iterations.

Participants reported a strong preference for online study due to flexibility and physical distance from peers, but noted challenges with distraction and inconsistent focus. Motivation was higher when studying with peers who shared similar goals, whereas studying with close friends sometimes reduced productivity, echoing prior findings on peer accountability and group dynamics (Berthelon et al., 2019; Zambrano et al., 2019). Trust, psychological safety, and clearly defined expectations were identified as critical conditions for effective co-study.

Feedback on the prototype indicated that structured study templates, timers, peer matching, and personal study statistics were essential features. Participants valued simplicity, low cognitive load, and clear progress indicators. Gamification was viewed positively when tied to meaningful progress rather than arbitrary rewards, consistent with prior gamification research in higher education (Schlosser et al., 2023; Smirani & Yamani, 2024).

Future work will involve testing refined prototypes with a more diverse graduate student population. Additional research is needed to evaluate effective peer-matching criteria, embedded learning strategies, and motivation systems tailored to graduate learners (Connell et al., 2023; Xu et al., 2021). If deployed, future iterations could integrate analytics to support data-driven design improvements and better understand long-term study behaviors.

References

- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. ?-?). <https://doi.org/10.59668/2109.21735>
- Thai, K.-P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. (2022). Learning engineering is human-centered. Dalam J. Goodell & J. Kolodner (Eds.), *Learning engineering toolkit: Evidence-based practices from the learning*

sciences, instructional design, and beyond (hlm. 84–123). Routledge. <https://doi.org/10.4324/9781003276579-7>

Totino, L., & Kessler, A. (2024). "Why did we do that?" A systematic approach to tracking decisions in learning engineering. International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE).

*) There are several other pieces of literature that we will include on the poster, related to collaborative study, intrinsic and extrinsic motivation, and also group size for the study connection platform.



Co-Designing AI-Enabled Learning in Nursing Education: A Learning Engineering Approach Using Faculty and Student Insights

Mamta Shah, Margaret Calacci, Lina Lander, & Celia Coochwyte

Academic-Industry Partnerships

Generative AI

Human-centered Design

learning engineering

Nursing Education

Technology Integration

Generative AI tools are rapidly entering nursing education. Yet little empirical work examines how faculty and students actually use these tools in authentic learning contexts or what design requirements emerge from their lived experiences. This practitioner-oriented study investigates an academic-industry partnership through the adoption of Sherpath AI (SPAI), a generative AI chatbot for nursing education, across 45 prelicensure course sections at Arizona State University's Edson College of Nursing and Health Innovation. Through semi-structured interviews with seven faculty members and focus groups with ten students, we identified patterns in AI appropriation for studying, clinical reasoning, and pedagogical design. Drawing on learning engineering (LE) frameworks emphasizing human-centered, iterative, and contextually grounded design, we surface key themes: AI as a cognitive and

pedagogical partner, trust calibration through evidence pathways, prompting as a metacognitive scaffold, and academic-industry learning loops. Findings reveal opportunities - personalized explanations, confidence building, and instructional coherence - and priorities, including verification pathways in specialized domains and scaffolded prompting support. We conclude with empirically derived design requirements for scalable, equitable, and trustworthy AI-enabled learning systems in safety-critical professional education.

Introduction

Generative AI has infiltrated nursing education at an unprecedented pace, promising to personalize learning, reduce faculty workload, and prepare students for technology-enabled practice. Yet, this rapid integration occurs within a high-stakes professional domain where conceptual, ethical, and procedural knowledge must converge to ensure safe, quality patient-centered care. Nurse educators face a precarious challenge: how to harness AI's potential for explanation, practice, and reasoning support while maintaining the rigorous standards required for clinical competence, professional judgment, and ethical practice.

This challenge is amplified by several factors unique to nursing education. First, the discipline requires integration across multiple knowledge areas, including pathophysiology, pharmacology, clinical procedures, ethical reasoning, and patient-centered communication— all of which must be applied in rapidly changing, high-pressure clinical contexts (Benner, Sutphen, Leonard, & Day, 2010). Second, nursing education operates under strict accreditation standards and licensure requirements, including the National Council Licensure Examination (NCLEX), which creates tension between pedagogical innovation and regulatory compliance. Third, the profession is experiencing severe faculty shortages and growing student enrollments, intensifying workload pressures while simultaneously demanding high-quality, individualized instruction (American Association of Colleges of Nursing, 2023).

Against this backdrop, generative AI tools offer both promise and peril. They can provide personalized cognitive scaffolding, generate practice materials aligned to learning objectives, and support self-directed study. However, they also raise concerns about accuracy in specialized clinical contexts, as well as overreliance that could undermine the development of clinical reasoning, academic integrity, and equal access. Most critically, there exists little empirical evidence about how nursing faculty and students use these tools in authentic learning contexts, the breakdowns they encounter, and what design features would support effective and safe integration.

Literature Review

Learning engineering (LE) provides a principled approach to addressing these challenges. Mishra (2025) frames LE as a human-centered, interdisciplinary, iterative, and values-driven practice that bridges learning sciences, instructional design, data analytics, and domain expertise. Rather than viewing technology adoption as the implementation of pre-designed tools, Kolodner (2023) suggested LE conceptualizes it as ongoing design-in-use, in which practitioners surface breakdowns and refinement opportunities through authentic activity.

Additionally, Saxberg (2017) emphasizes that learning innovations succeed only when grounded in cognitive task design; that is, understanding the specific knowledge structures, reasoning patterns, and performance demands of a domain. For nursing, this means attending to how students develop clinical reasoning (hypothesis generation, evidence evaluation, decision-making under uncertainty), integrate theoretical knowledge in clinical situations, and form professional identity. Goodell, Kolodner, and Kessler (2022) extend this by arguing that LE requires iterative cycles of design → implementation → data collection → refinement, with close collaboration among educators, learners, researchers, and developers.

Mishra, Warr, and Islam (2023) describe generative AI as protean (taking many forms), opaque (lacking transparency in reasoning), unstable (outputs varying across iterations), and social (shaped by and shaping human interaction). These characteristics demand that educators develop what they term a "teaching compass" or competencies: agency in shaping AI use, attention to well-being and workload sustainability, and knowledge, skills, values, and ethics for navigating AI-enabled systems (Mishra & Oster, 2023). These framing positions educators not as passive adopters but as active designers who must integrate technical, pedagogical, and contextual knowledge across learning strategies.

Recent systematic reviews provide mixed evidence on AI's impact on nursing education. Ma et al. (2025) analyzed 47 studies and found that AI tools can enhance knowledge acquisition, self-directed learning, positive attitudes toward technology, and simulation performance. However, the effects on clinical reasoning and critical thinking were inconsistent. Similarly, Ramírez-Baraldes and colleagues (2025) identified opportunities for AI to support personalized learning pathways and reduce faculty administrative burden, while highlighting the risks related to accuracy, bias, academic integrity, and overreliance on the output.

Notably, most existing research examines AI-enabled simulations, intelligent tutoring systems, or predictive analytics for student success, not generative AI chatbots for open-ended learning support. The handful of studies on generative AI in nursing education (predominantly focused on ChatGPT) reveal faculty concerns about accuracy in specialized domains like pediatric dosing calculations, potential for students to bypass critical thinking. Moreover, by uncritically accepting AI-generated outputs, educators may struggle to determine whether students have developed true conceptual understanding versus only surface-level familiarity (Chan & Lee, 2024; Phillips & Lee, 2024).

Crucially, existing literature lacks empirical investigation on how nursing faculty and students appropriate generative AI tools in authentic course contexts, what strategies they develop, how they calibrate trust, and what design features would support rather than undermine professional competence development. This gap is particularly significant given nursing education's emphasis on situated cognition, the principle that knowledge must be developed and demonstrated in contexts approximating clinical practice (Lave & Wenger, 1991).

Research Questions

This study addresses these gaps by examining early adoption of Sherpath AI (SPAI), Elsevier's generative AI chatbot for nursing education, at Arizona State University's Edson College of Nursing and Health Innovation. We ask:

1. How do nursing students and faculty use generative AI in authentic learning contexts across diverse course types?
2. What opportunities, constraints, and risks do they identify through their lived experiences?
3. How can their insights inform learning-engineered design requirements for future AI-enabled learning systems in professional education?

By positioning faculty and students as co-designers whose everyday use reveals design requirements, we demonstrate how learning engineering principles translates into specific features for AI-enabled learning systems in safety-critical domains.

Methods

Context and Participants

Arizona State University's Edson College of Nursing and Health Innovation, a large public university serving a diverse student body, adopted Sherpath AI (SPAI) in Summer 2024 as an optional learning resource. SPAI is a generative AI chatbot designed for nursing education, featuring evidence-linked citations from Elsevier's nursing textbooks (Lewis's Medical-Surgical Nursing, Hockenberry's Wong's Nursing Care of Infants and Children), competency alignment to nursing curricula, and transparent citations to authoritative sources. SPAI was available to all participants via their course learning management system (Canvas). In-site (just-in-time) onboarding and on-demand support were provided by co-authors at ASU. Data were collected from nursing faculty (N = 7) and students (N = 10) who used SPAI across 45 prelicensure nursing course sections (Summer 2024–Fall 2025). Faculty participants taught courses such as pediatric nursing, critical care, nursing research foundations, and experiential/clinical coordination. Student participants spanned second through fourth year with varying engagement levels and prior AI experience.

Data Collection and Analysis

Following IRB approval (#00022071), the research team conducted semi-structured faculty interviews (individual, 45-60 minutes) and student focus groups (3 groups of 3-4 students, 60 minutes) between Fall 2024 and Fall 2025. Semi-structured protocols allowed flexibility to pursue emergent themes while ensuring systematic coverage of key domains: (1) awareness and implementation; i.e., how participants described and integrated SPAI across courses; (2) use patterns; i.e., types of prompts, workflows, and refinement strategies; (3) impact; i.e., specific examples of learning/teaching support; (4) trust and quality; i.e., evaluation of accuracy, verification practices, and factors influencing confidence; (5) barriers and improvements; i.e., challenges encountered and desired features; and (6) additive value; i.e., comparison to traditional resources and influence on practice. Sessions were audio-recorded, transcribed verbatim, and member-checked. We employed thematic analysis (Braun & Clarke, 2006) to identify patterns in SPAI appropriation, trust reasoning, prompting strategies, and challenges. Four themes emerged: (1) AI as a cognitive and pedagogical partner, (2) trust calibration and evidence pathways, (3) prompting as a learning engineering process, and (4) academic-industry learning loops. Analysis attended to 'breakdowns' revealing design requirements (Kolodner, 2023) and 'teaching compass' competencies (Mishra & Oster, 2023).

Results

AI as a Cognitive and Pedagogical Partner

Students consistently used SPAI to deconstruct complex concepts, compare diagnoses, and rehearse NCLEX-style reasoning. One student explained, "Sometimes I just need it to explain something at my level—like third semester—not a whole textbook chapter." This personalization was particularly valued in courses where textbooks provided comprehensive but overwhelming detail. Another student noted that SPAI helped build confidence before class participation: "I don't want to ask a dumb question in front of everyone. I [will] try it in Sherpath AI first." Faculty appropriated SPAI for pedagogical design. They used it to generate case studies, create NCLEX-style practice questions with appropriate distractors, and scaffold active learning activities. One instructor shared, "I used it to check that my explanations matched what's in the textbook and what students would see on their exams." SPAI also supported pedagogical coherence across instructors by offering standardized, evidence-linked explanations.

Trust Calibration, Accuracy, and Evidence Pathways

Participants consistently emphasized SPAI's evidence-linked citations as the primary source of trust, distinguishing it from general-purpose AI tools. A faculty member noted, "If I see Lewis or Hockenberry cited, I know where it's coming from." Students echoed this sentiment: "I check the evidence tab every time. ChatGPT never tells me where it gets anything."

This transparency enabled what we term "trust calibration," where participants developed nuanced judgments about when to rely on SPAI and when additional verification was needed. Yet, faculty also identified limits, especially in pediatrics: "Sometimes it gets the nuance wrong, so I double-check against clinical practice." These tensions reflect nursing education's need for clinically aligned, continuously updated content, aligned with Mishra's principle (2025) that AI systems inevitably embed trade-offs requiring human judgment and oversight.

Prompting as a Learning Engineering Process

Early use revealed prompting challenges: unclear requests, overly broad queries, and outputs mismatched to the learner's level. Through experimentation, students and faculty learned to refine prompts: "Explain this again but simpler," "Give me distractors a second-semester student would choose," "Create a two-part case study with increasing complexity." This iterative prompting became a metacognitive scaffold. As one student admitted, "I realized if I can't ask a good question, it's because I don't understand what I'm stuck on." Such iterative refinement exemplifies LE cycles of design → feedback → redesign, and points to design needs such as built-in level controls, clarifying questions, suggested follow-up questions, and competency alignment.

Academic–Industry Learning Loops

Faculty identified additional needs: multimodal learning assets (audio summaries, diagrams, side-by-side comparisons), competency-aligned question generation, instructor insight filters, and prompt exemplars. Notably, SPAI functioned as what Star and Griesemer (1989) term a "boundary object," i.e., a shared referent that faculty, clinical partners, and students used to triangulate conceptual understanding and align instruction across classroom and clinical settings. One instructor noted, "I take the Sherpath AI-generated STEMI summary to my ICU colleagues and ask, 'Is this what you're seeing clinically?'" These patterns reflect Kolodner's (2023) claim that LE is fundamentally collaborative and that effective systems emerge from distributed expertise, not tool adoption alone.

Discussion and Conclusions

Sherpath AI adoption at ASU unfolded as a LE process: faculty and students iteratively adapted the tool, reflected on mismatches, and reworked instructional strategies. This aligns with Mishra, Warr, and Islam's (2023) arguments that educators must integrate technical, pedagogical, and contextual knowledge within AI-mediated learning systems. It also resonates with Kolodner's (2023) framing of LE as design-in-use, where authentic activity reveals breakdowns that guide redesign. Findings also mirror nursing literature: AI can enhance self-directed learning, confidence, and foundational knowledge, but requires safeguards to support clinical reasoning, ethical judgment, and professional identity formation (Ma et al., 2025). The ASU–Elsevier partnership demonstrates how academic–industry collaboration can enable principled, data-informed, human-centered innovation. As Saxberg (2017) notes, scaling effective learning requires structures for exposure, education, effort, and evaluation. These structures emerged organically through this partnership.

Three insights emerged from this work. First, generative AI tools serve different functions for different users in different contexts: students sought cognitive scaffolding while faculty sought pedagogical design support. Second, trust in AI is not binary but contextual and calibrated, requiring transparent evidence pathways and verification opportunities. Third, knowing what to ask, when to use the tool, and how it supports learning is determined through practice and highlights iterative design needs for scaffolded interfaces that support users at varying expertise levels.

This practitioner-oriented study also has several takeaways for specific audiences. For nurse educators and programs, GenAI can be positioned as a cognitive partner that supports explanation, comparison, and rehearsal without replacing human judgment. Embedding AI literacy (prompting, verification, ethical use) into curricula strengthens both professional formation

and safe clinical reasoning. For academic institutions, AI adoption can function as an infrastructure for learning engineering, where faculty and student feedback regularly informs tool refinement. Intentional structures such as policies, training, and course-level experimentation are needed to support effective and ethical GenAI use in professional programs. For industry partners, faculty, and students act as co-designers and end-users. Academic–industry partnerships should be structured as ongoing learning loops feeding authentic insight into product evolution. For learning engineers, nursing education demonstrates how LE principles such as cognitive alignment, iterative design, contextual constraints, and human-centered values translate into specific AI features such as level controls, competency tagging, and trust signals. This context underscores the need for LE teams to integrate learning sciences, domain expertise, data, and ethics in safety-critical fields.

References

- American Association of Colleges of Nursing. (2023). Fact sheet: Nursing faculty shortage. <https://www.aacnnursing.org/news-data/fact-sheets>
- Benner, P., Sutphen, M., Leonard, V., & Day, L. (2010). Educating nurses: A call for radical transformation. Jossey-Bass.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.
- Chan, C. K. Y., & Lee, K. K. W. (2024). Generative AI in nursing education: Opportunities and challenges. *Nurse Education Today*, 134, 106089.
- Goodell, J., Kolodner, J., & Kessler, A. (2022). Learning engineering applies the learning sciences. In *Learning engineering toolkit* (pp. 47-82). Routledge.
- Kolodner, J. L. (2023). Learning engineering: What it is, why I'm involved, and why I think more of you should be. *Journal of the Learning Sciences*, 32(2), 305-323.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.
- Ma, J., Wen, J., Qiu, Y., Wang, Y., Xiao, Q., Liu, T., ... & Sun, Z. (2025). The role of artificial intelligence in shaping nursing education: A comprehensive systematic review. *Nurse Education in Practice*, 104345.
- Mishra, P. (May, 2025). From artifacts to systems: The evolution of learning engineering. Keynote. International Education Forum on Learning Engineering. Ankara, Turkey.
- Mishra, P., Warr, M., & Islam, R. (2023). TPACK in the age of ChatGPT and Generative AI. *Journal of Digital Learning in Teacher Education*, 39(4), 235-251.
- Phillips, A., & Lee, S. (2024). ChatGPT in nursing education: Faculty perspectives on risks and opportunities. *Journal of Nursing Education*, 63(3), 156-162.

Ramírez-Baraldes, E. L., García-Gutiérrez, D., & García-Salido, C. (2025). Artificial intelligence in nursing: New opportunities and challenges. *European Journal of Education*, 60(1), e70033.

Saxberg, B. (2017, April). Learning engineering: the art of applying learning science at scale. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale* (pp. 1-1).

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3), 387-420.



Developing a Model to Support Collaborative Engineering Projects: Student Uncertainty as a Productive Resource for Engineering (SUPER-E)

Ye, Y, Jordan, M. E., & Chen, Y.C.

Collaboration

Engineering Learning

K-12

Student Uncertainty

Engineering design in K–12 classrooms is inherently uncertain as students tackle open-ended problems and team dynamics. Building on the Student Uncertainty as a Pedagogical Resource (SUPeR) model in science, informed by engineering design education theory and teachers’ design activity and discussion in a professional development, we developed the Student Uncertainty as a Productive Resource for Engineering (SUPER-E) model. SUPER-E is a four-phase framework that maps uncertainty across the engineering design process. It supports teachers to anticipate when and how uncertainties emerge across each phase and offers teachers a structured yet flexible tool for leveraging student uncertainty as a productive resource. Next steps involve classroom-based implementation with middle school students.

Introduction

Engineering design is an inherently ill-structured, uncertainty-filled activity in K–12 settings. Students must manage hesitancy, differing opinions, and conflict while collaborating in discourse-intense, argumentation-rich engineering cultures (Cunningham & Kelly, 2017; Jordan, 2015). How teachers design and facilitate engineering projects therefore shapes whether students can notice, acknowledge, and productively work with their uncertainty. Previous work has primarily examined how students navigate scientific uncertainty in science classrooms. However, because science seeks to explain how the world works, whereas engineering focuses on designing scalable solutions to real-world problems using science as one tool in that endeavor (Goodell et al., 2023), there is a clear need for a distinct model tailored to uncertainty in engineering design contexts.

Developing the SUPER-E Model

Initial Development: Theory Informed

To develop the Student Uncertainty as a Productive Resource for Engineering (SUPER-E) model, we initially identified the sources of uncertainty in collaborative engineering design projects to create an initial tentative model. This work was informed by three resources: previous research on uncertainty in engineering design, our previous work on the (Student Uncertainty as a pedagogical resource (SUPeR) model, and our own experience facilitating engineering design projects with K-12 teachers and students. Across these, we consider two aspects of uncertainty as important: the types of uncertainty that students are likely to experience (i.e., conceptual, epistemic, and relational), and (b) how teachers can help students navigate those uncertainties (e.g., Kaur & Dasgupta, 2024).

The SUPER-E model was also informed by understanding of how people navigate uncertainty, particularly how teachers can help students navigate uncertainty (Chen & Jordan., 2024). Uncertainty navigation can be understood as the process of strategically raising, maintaining, reducing, and ignoring different types of uncertainty in order to advance learning in engineering design contexts (Jordan & McDaniel, 2014; Chen et al., 2024). Raising uncertainty requires teachers to know what desirable uncertainties we need to identify or bring up so they can be addressed. Maintaining uncertainty means encouraging students to “live with” desirable uncertainties for a while because it is unproductive to reduce them too soon. Reducing uncertainty means what uncertainty we want to get rid of because it is undesirable. It is also necessary to know when and how to help students postpone or ignore uncertainties that are irrelevant to the task at hand.

Iterating the Model: Teacher Informed

After developing an initial SUPER-E model (Figure 1), we tested it with 18 in-service science teachers during a week-long PD on using uncertainty as a resource for STEM learning. Teachers worked in teams on a Solar Home Hack engineering design challenge organized around the first and second versions of the model, with their work audio–video recorded. Their activity was punctuated by periodic whole group discussion, reflection, and critique of the model. In this way, teachers stepped into the design challenge as students and stepped out as expert educators to offer critique and to reflect on their own experiences of uncertainty during the engineering design process. In a final discussion, teachers identified types and sources of uncertainty students might encounter at each phase and recommended further differentiating Phases 2 and 3 into sub-sections, leading to a revised SUPER-E model (Figure 2).

Figure 1.

Initial SUPER-E Model

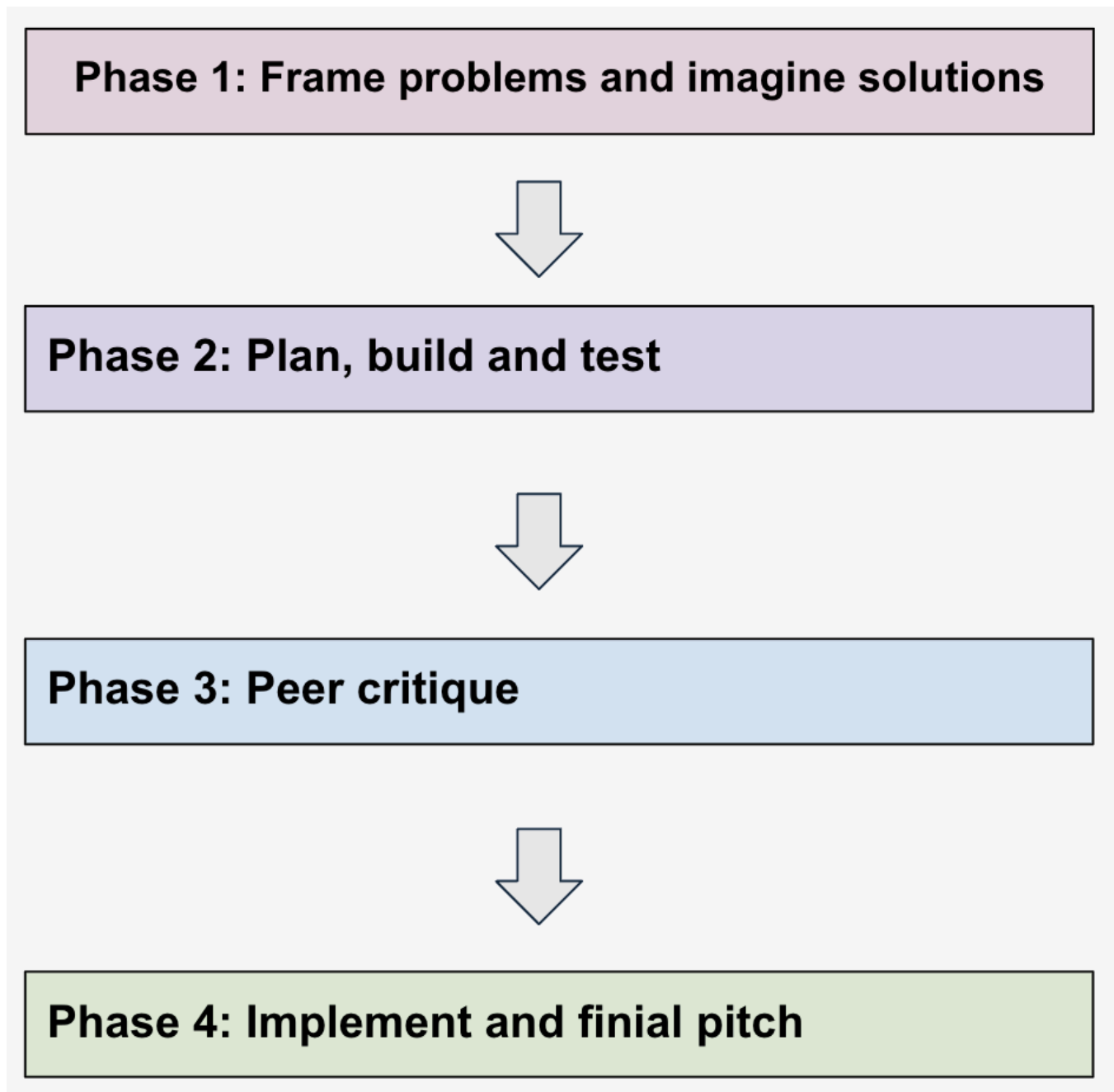
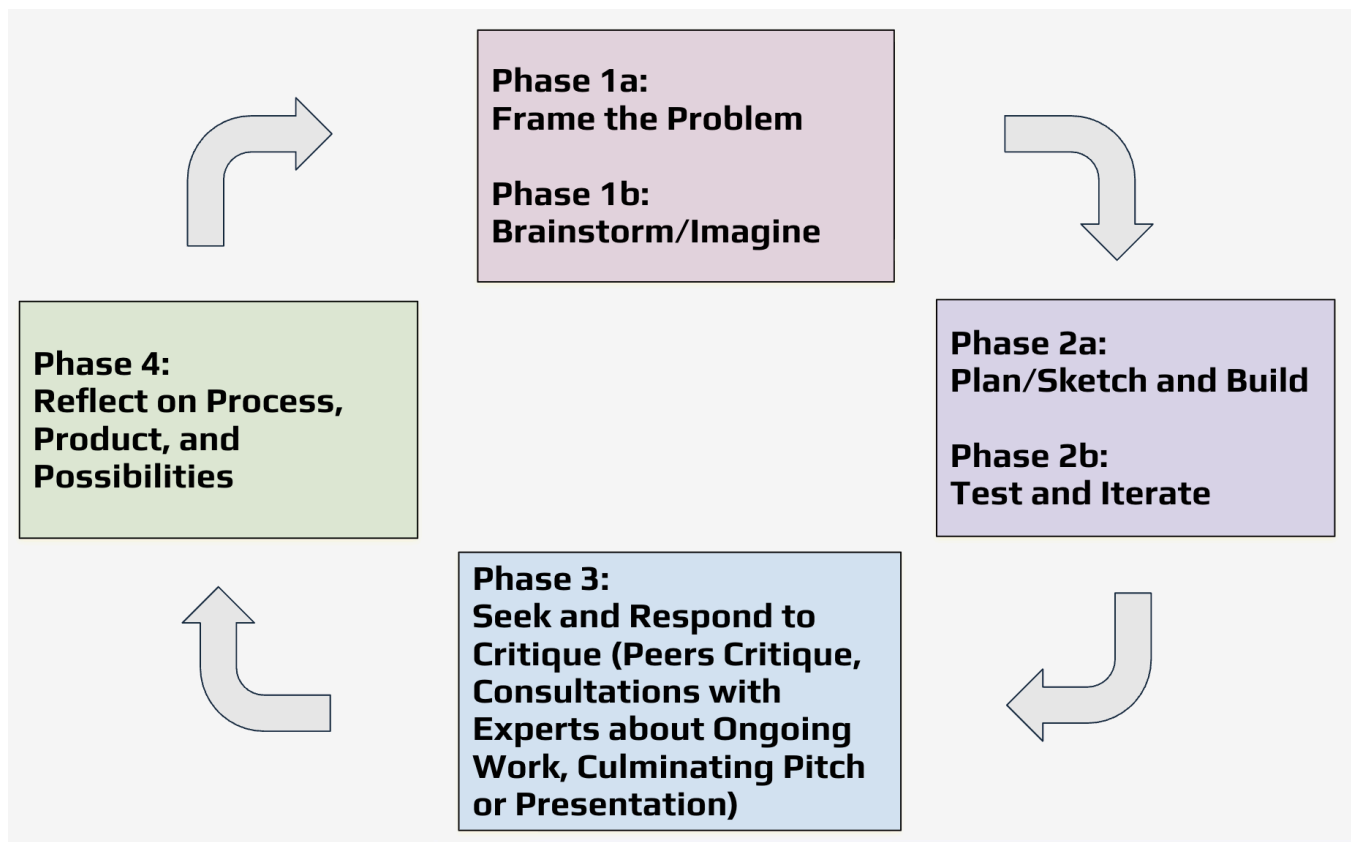


Figure 2.

Iterated SUPER-E Mode



The SUPER-E Model

The iterated SUPER-E model consists of four phases as described below. During each phase, particular types and sources of uncertainty are likely to arise for students, which we categorize as conceptual, epistemic, and relational. Teachers need to be aware of and plan for uncertainties likely to arise at each phase so they can support students' uncertainty navigation in how they create engineering design challenges and respond to student uncertainties throughout the engineering design process.

Phase 1 consists of two sub-stages. In 1a—Frame the Problem, students identify initial criteria and constraints while the teacher contextualizes the challenge, and students' uncertainties and curiosities emerge. In 1b—Brainstorm/Imagine, students generate and evaluate multiple ideas and consider integration with existing systems, deliberately raising conceptual and epistemic uncertainty to clarify the problem space, goals, and constraints, justify why ideas might work, compare competing approaches, and sustain multiple possibilities for deeper reasoning and richer idea generation—even if the group's thinking is temporarily incoherent. Making trade-offs between criteria and constraints (Kelly & Cunningham, 2019) helps reduce relational uncertainty by establishing shared criteria for evaluating, refining, and integrating solutions.

Phase 2 has two sub-stages: 2a—Plan/Sketch and Build and 2b—Test and Iterate, where students turn plans into tangible products and collect, interpret, and curate results. In this hands-on "conversation with materials" (Bamberger & Schön, 1983), conceptual uncertainty is common and sometimes must be reduced (e.g., confusion about essential requirements) to keep work moving, while epistemic uncertainty is best maintained to support iterative reasoning about design decisions and outcomes. Clarifying team roles helps reduce relational uncertainty, promote equitable participation, and minimize unproductive tensions during building.

Phase 3-Seek and Respond to Critique involves peer critique, expert consultation, and a culminating pitch or presentation, with students deliberately seeking and responding to feedback on their ongoing work. Some epistemic and relational uncertainties are intentionally maintained—for example, “living with” conflicting critiques or differing perspectives within teams or with experts. Ultimately, students’ uncertainty will be reduced by reconciling critique, converging on feasible solutions, and presenting a defensible design in a final pitch that establishes accountability and justification to peers, clients, or community stakeholders.

Phase 4-Reflecting on Product, Process, and Possibilities engages students in structured reflection to consolidate what they learned and extend their thinking. They identify what remains uncertain about their designs, consider future iterations, and recognize design as ongoing and iterative. Students explore new uncertainties (e.g., installation, usability, social embeddedness), reflect on how they collaborated and responded to uncertainty, and examine broader social, economic, and environmental implications, including tradeoffs and ambiguities. Because engineering design always entails tradeoffs (Kelly & Cunningham, 2019), this phase offers space to grapple with ambiguity tied to multiple, or conflicting values and to reframe uncertainty as a productive driver of deeper reasoning, and systems-level thinking.

Conclusion

The SUPER-E model offers a structured yet flexible framework for helping K–12 students use uncertainty as a productive resource in engineering design. By addressing conceptual, epistemic, and relational uncertainties across four phases and clarifying how teachers can support students, it shows how collaboration, critique, and reflection deepen engagement and make design work more authentic. The model not only maps the iterative nature of engineering but also foregrounds often overlooked affective and social dimensions of learning.

References

- Bamberger, J., & Schön, D. A. (1983). Learning as reflective conversation with materials: Notes from work in progress. *Art Education*, 36(2), 68–73. <https://doi-org.ezproxy1.lib.asu.edu/10.1080/00043125.1983.11653404>
- Chen, Y. C., & Jordan, M. (2024). Student uncertainty as a pedagogical resource (SUPeR) approach for developing a new era of science literacy: Practicing and thinking like a scientist. *Science Activities*, 61(1), 1-15.
- Chen, Y. C., Jordan, M., Park, J., & Starrett, E. (2024). Navigating student uncertainty for productive struggle: Establishing the importance for and distinguishing types, sources, and desirability of scientific uncertainties. *Science Education*, 108(4), 1099-1133.
- Cunningham, C. M., & Kelly, G. J. (2017). Framing engineering practices in elementary school classrooms. *International Journal of Engineering Education*, 33(1), 295–307.
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning Engineering at a Glance. *Journal of Military Learning*, 7(1)
- Jordan, M. E. (2015). Variation in students’ propensities for managing uncertainty. *Learning and Individual differences*, 38, 99-106.
- Jordan, M. E., & McDaniel Jr, R. R. (2014). Managing uncertainty during collaborative problem solving in elementary school teams: The role of peer influence in robotics engineering activity. *Journal of the Learning Sciences*, 23(4), 490-536.
- Kaur, N., & Dasgupta, C. (2024). Investigating the interplay of epistemological and positional framing during collaborative uncertainty management. *Journal of the Learning Sciences*, 33(1), 80–124.

<https://doi.org/10.1080/10508406.2024.2312163>

Kelly, G. J., & Cunningham, C. M. (2019). Epistemic tools in engineering design for K-12 education. *Science Education*, 103(4), 1080-1111.



Developing Learning Strategy Heuristics for Active Mobile Learning Platforms

Ishrat Ahmed, Shailee Shah, & Scotty D. Craig

Active Learning

learning engineering

Learning Strategy Analysis

Mobile Learning

Usability Heuristics

This study presents a heuristic framework and four-level rubric for evaluating the pedagogical effectiveness of mobile learning applications. Drawing on literature in active learning and mobile learning platforms, we developed an eight-category framework encompassing active learning, audio-based engagement, content access, interactivity, collaboration, digital equity, instructor feasibility, and feedback. Each category is operationalized through sub-heuristics and assessed using a rubric that rates implementation quality from no implementation to full pedagogical integration. Situated within the Learning Engineering (LE) process, this work presents a nested design cycle within the Challenge phase, in which a heuristic checklist was created to systematically identify pedagogical strengths and gaps in existing mobile learning platforms. Using a multi-phase methodology, literature review, and framework synthesis and validation, we show how this approach offers actionable, evidence-based criteria for identifying strengths and gaps in mobile learning platforms. This framework is designed to support educators and developers in creating mobile learning environments that move beyond

content delivery to foster deeper engagement, learner autonomy, and more effective learning outcomes.

Introduction

Advances in educational technology have reshaped how students in higher education access, and interact with learning content. Yet this shift also presents a challenge as learning environments become increasingly digital and self-directed, ensuring that students remain actively engaged. Within Active L@S, this challenge underscores the need for instructional approaches that promote agency, and sustained interaction with course material. Mobile learning is one such method, offering a scalable way to extend learning beyond traditional classroom boundaries. Mobile learning enables instruction that is flexible, self-paced, and responsive to diverse learner needs, reflecting a broader shift from traditional, in-person models toward technology-enhanced education. While mobile platforms expand access and convenience, their pedagogical impact depends on how well they support meaningful learning. Active learning, defined by student engagement, reflection, and learner agency, has emerged as a critical benchmark because it fosters deeper cognitive processing and improved learning outcomes (Miyatsu, Nguyen, & McDaniel, 2018; Roscoe & Craig, 2022). This makes it essential to evaluate whether mobile learning tools are strategically supporting learners, rather than just delivering content.

Despite the growing adoption of mobile learning tools, structured and research-based approaches for evaluating their pedagogical alignment remain limited. Existing frameworks tend to prioritize usability, access, and content delivery while overlooking the deeper cognitive and behavioral processes that support sustained learning. As a result, many mobile learning tools remain supplementary rather than strategically integrated into instruction (Rangel-de Lazaro and Duarte, 2023) and common barriers, such as insufficient instructional design and lack of alignment with evidence-based strategies continue to limit their effectiveness (Sophonhiranrak, 2021). Foundational models, such as Ozdamli and Cavus's (2011) framework for mobile learning elements, highlight the importance of harmonizing learner needs, content delivery, instructional roles, learning environments, and assessment, yet no widely adopted evaluation tool has operationalized these principles for practical use by educators and designers.

Learning Strategy Analysis (LSA) addresses this gap by providing a structured, research-informed methodology for assessing the pedagogical alignment of digital learning systems with evidence-based instructional strategies (Villa, Craig, Zakhidov, & Zielke, 2021). Situated within the Learning Engineering (LE) process, LSA was employed in this study as a nested analytic cycle (Craig et al., 2025) within the Challenge phase to inform the design of a heuristic checklist applicable to mobile learning platforms. Specifically, LSA guided the identification and operationalization of instructional strategies such as elaboration, self-explanation, multimedia practice, or spaced practice, strategies known to foster deeper engagement and mastery. Using LSA-informed heuristics, checklists, and rubrics, the resulting framework enables systematic identification of pedagogical strengths and gaps in mobile learning applications, supporting evidence-based decision-making, and learner-centered design. By integrating LE and LSA, this approach bridges the gap between technological deployment and pedagogical integrity, providing a practical tool for evaluating and improving the instructional quality of mobile learning environments.

Methodology

This study was conducted collaboratively by two researchers and employed a two-phase methodology to develop a pedagogically grounded framework for evaluating mobile learning applications. Guided by principles of active learning and

evidence-based instructional design, the research aimed to create a practical heuristic tool that educators and designers can use to assess whether mobile platforms meaningfully support learner engagement, cognitive processing, and instructional effectiveness.

The methodology involved two phases. Phase 1 involved a targeted literature review on mobile learning and active learning strategies to identify relevant pedagogical constructs (e.g., elaboration, note-taking, and multimedia learning) and establish foundational design principles. Phase 2 synthesized these insights into an eight-category heuristic framework, covering areas such as Active Learning, Learning with Voices/Audio, Content Access, Engagement and Interactivity, Communication and Collaboration, Digital Divide, Instructor Readiness and Feasibility, and Feedback. Each category was further divided into sub-heuristics that include actionable design recommendations. These heuristics function both as evaluative criteria for identifying areas of improvement and as guidelines for designing pedagogically grounded mobile learning tools. While not all strategies must be present in a given system, the inclusion of any single sub-heuristic, or a meaningful combination of them, can enhance active learning.

Learning Strategy Heuristics Related to Mobile Learning Platforms

As outlined above, we developed eight heuristic categories for evaluating mobile learning systems. The following section provides definitions of each heuristic and describes its corresponding sub-heuristic categories.

Active Learning Heuristic

Active learning refers to instructional methods that engage learners in the process of knowledge construction through different activities. The Active Learning heuristic emphasizes subheuristics that prompt learners to construct knowledge, including note-taking, self-explanation, summarization, and question-answering (Miyatsu, Nguyen, & McDaniel, 2018; Roscoe & Craig, 2022). Mobile systems should support structured, paraphrased note-taking (Van Wyk & Van Ryneveld, 2018); prompt learners to explain concepts in their own words beyond multiple-choice (Bisra et al., 2018); enable meaningful summarization rather than copy-paste behavior; and incorporate reflective, open-ended questions that promote elaborative reasoning. Questions that prompt elaborative reasoning are more effective in assessing student understanding than factual or binary questions (Hernández-de-Menéndez et al., 2019). These strategies collectively foster deep cognitive processing, sustained engagement, and improved learning outcomes.

Learning with Voices/Audio Heuristic

This heuristic emphasizes the pedagogical value of audio-based learning in mobile contexts. Effective systems should incorporate sub-heuristics such as delivering high-quality audio, providing playback controls (e.g., pause, rewind, and timestamp tagging) that support review and note-taking without disrupting the listening experience (Schroeder et al., 2020), and integrating speech-to-text features for accessibility needs (Craig & Shroeder, 2019). These elements make audio a powerful tool for mobile learning environments.

Content Access Heuristic

This heuristic refers to how easily learners can access both platform-provided content and their own user-generated content (e.g., notes, highlights). Mobile systems should offer sub-heuristics such as seamless synchronization across devices (Kiewra et al., 1991), timestamp-linked navigation between notes and content (Kiewra et al., 1991), and personal archives that are searchable and easy to review (Kiewra et al., 1991). Such features support autonomy, revisitation, and reflection, key components of self-regulated learning.

Engagement & Interactivity

Heuristic This heuristic focuses on learner engagement (i.e., motivational and behavioral involvement in learning) and interactivity (i.e., learner's ability to manipulate content or receive feedback from the system). Interactive elements, when grounded in pedagogy, can help maintain motivation and foster persistence. Gamification elements such as badges or points can enhance motivation (Sophonhiranrak, 2021), while opportunities for simulations or applied tasks help learners connect knowledge to authentic contexts (Rangel-de Lazaro & Duarte, 2023; Ozdamli & Cavus, 2011). Together, these features can sustain attention in mobile environments where distractions are common.

Communication & Collaboration Heuristic

This heuristic assesses the system's capacity to support interpersonal learning, peer exchange, and instructor-learner interaction. Such interaction reinforces understanding, supports metacognition, and builds learning communities. Effective platforms enable peer interaction through shared notes, discussion spaces, or review tools, and support instructor communication through messaging or comments, and finally integrations with third-party tools like WhatsApp or YouTube expand opportunities for interaction, fostering learning communities and deeper understanding (Sophonhiranrak, 2021; Rangel-de Lazaro & Duarte, 2023; Ozdamli & Cavus, 2011).

Digital Divide Heuristic

This heuristic examines the platform's capacity to accommodate diverse learners, particularly those with limited digital resources. Systems should function smoothly across a range of devices and operating systems (Sophonhiranrak, 2021; Ozdamli & Cavus, 2011), offer offline functionality for core features, and provide multilingual support (Sophonhiranrak, 2021; Ozdamli & Cavus, 2011). These elements help ensure that mobile learning reduces rather than reinforces inequities.

Instructor Readiness & Feasibility

This heuristic evaluates how easily instructors can adopt and integrate the platform into their teaching practice. Effective systems provide intuitive onboarding resources (Sophonhiranrak, 2021; Rangel-de Lazaro & Duarte, 2023), support flexible content authoring, and ensure compatibility to integrate with common learning management systems (Ozdamli & Cavus, 2011).

Feedback

Feedback refers to the quality and timeliness of information provided to learners about their performance and engagement. Mobile systems should offer appropriately timed feedback, clear and constructive explanations (Ruth et al, 2021; Shute, 2008), and options for learners to choose the depth and frequency of feedback they receive. When given a choice, learners are more likely to view feedback and show more positive engagement with the system (Kuklick, 2025). Support for dialogue-based feedback further enhances engagement and reflection. Together, these features promote self-regulation, deeper learning, and sustained use.

Learning Strategy Heuristics Rubric

Following the heuristics described above, we created a four-level rubric for each sub-heuristic. The four levels are: No Implementation (0), indicating the feature is entirely absent; Partial Implementation (1), indicating minimal or incomplete presence; Adequate Implementation (2), indicating functional but not fully optimized support; and Full Implementation (4), indicating comprehensive, well-integrated pedagogical alignment. For example, the "prompts for self-explanation" sub-heuristic in Active Learning heuristic ranges from no prompts for learners to articulate understanding, to occasional generic

prompts, to regular but non-adaptive prompts, and finally to fully embedded, context-aware prompts that encourage deep elaboration. For a full list of rubrics, check [here](#).

The rubric offers actionable metrics with clear criteria to assess how well an app supports learner engagement, accessibility, and instructional integration. A total score out of 28 indicates overall alignment, with higher scores reflecting stronger support for effective, equitable, and interactive mobile learning experiences.

Conclusion

This study presents a structured, pedagogically grounded heuristic framework for evaluating mobile learning applications. By integrating evidence-based instructional principles with a four-level evaluative rubric, ranging from no implementation to full implementation, the framework enables systematic assessment of both the presence and quality of learning-supportive features. The application of this framework can demonstrate its practical value as both a diagnostic tool and a design guide, helping educators and developers identify strengths, address gaps, and enhance the instructional integrity of mobile platforms. Ultimately, this approach supports the creation of mobile learning systems that move beyond content delivery to foster deeper engagement, learner autonomy, and meaningful, high-impact learning.

Acknowledgement

The research reported here was partially supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

- Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, 30(3), 703–725. <https://doi.org/10.1007/s10648-018-9434-x>
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale* (pp. 115-132). <https://doi.org/10.59668/2109.21735>
- Carless, D., & Winstone, N. (2023). Dialogic feedback in higher education: A systematic review and conceptual framework. *Studies in Educational Evaluation*, 78, 101276. <https://doi.org/10.1016/j.stueduc.2023.101276>
- Craig, S. D., & Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Computers & Education*, 114, 193–205. <https://doi.org/10.1016/j.compedu.2017.07.003>
- Hernández-de-Menéndez, M., Vallejo Guevara, A., Tudón Martínez, J. C., Hernández Alcántara, D., & Morales-Menéndez, R. (2019). Active learning in engineering education: A review of fundamentals, best practices and experiences. *International Journal on Interactive Design and Manufacturing*, 13(3), 909–922. <https://doi.org/10.1007/s12008-019-00557-8>
- Kiewra, K. A., Benton, S. L., Kim, S. I., Risch, N., & Christensen, M. (1995). Effects of note-taking format and study technique on recall and relational performance. *Contemporary Educational Psychology*, 20(2), 172–187. <https://doi.org/10.1006/ceps.1995.1011>

- Kuklick, L. (2025). Effects of learner choice over automated, immediate feedback. *Learning and Instruction*, 96, Article 102065. <https://doi.org/10.1016/j.learninstruc.2024.102065>
- Mari van Wyk, & Linda van Ryneveld. (2018). Affordances of mobile devices and note-taking apps to support cognitively demanding note-taking. *Education and Information Technologies*, 23(4), 1639–1653. <https://doi.org/10.1007/s10639-017-9684-0>
- Miyatsu, T., Nguyen, K., & McDaniel, M. A. (2018). Five popular study strategies: Their pitfalls and optimal implementations. *Perspectives on Psychological Science*, 13(3), 390–407. <https://doi.org/10.1177/1745691617710510>
- Mullet, H. G., Butler, A. C., Berdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition*, 3(4), 222–229. <https://doi.org/10.1016/j.jarmac.2014.05.001>
- Ozdamli, F., & Cavus, N. (2011). Basic elements and characteristics of mobile learning. *Procedia – Social and Behavioral Sciences*, 28, 937–942. <https://doi.org/10.1016/j.sbspro.2011.11.173>
- Rangel-de Lázaro, G., & Duarte, J. M. (2023). Moving learning: A systematic review of mobile learning applications for online higher education. *Journal of New Approaches in Educational Research*, 12(2), 198–224. <https://doi.org/10.7821/naer.2023.7.1287>
- Roscoe, R. D., & Craig, S. D. (2022). Learning through collaborative explanation: Processes and outcomes. *Educational Psychologist*, 57(3), 191–206.
- Rüth, M., Breuer, J., Zimmermann, D., & Kaspar, K. (2021). The effects of different feedback types on learning with mobile quiz apps. *Frontiers in Psychology*, 12, 665144. <https://doi.org/10.3389/fpsyg.2021.665144>
- Schroeder, N. L., Glahn, R., & Craig, S. D. (2020). The effects of pedagogical agent voice and animation on learning, motivation, and cognitive load. *Computers & Education*, 149, 103814.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189. <https://doi.org/10.3102/0034654307313795>
- Sophonhiranrak, S. (2021). Features, barriers, and influencing factors of mobile learning in higher education: A systematic review. *Heliyon*, 7(4), e06696. <https://doi.org/10.1016/j.heliyon.2021.e06696>
- Van Wyk, M., & Van Ryneveld, L. (2018). Affordances of mobile devices and note-taking apps to support cognitively demanding note-taking. *Education and Information Technologies*, 23(4), 1639–1653. <https://doi.org/10.1007/s10639-017-9684-0>
- Villa, A., Craig, S. D., Zakhidov, A., & Zielke, M. (2021). Evaluating digital learning environments with the Learning Strategy Analysis framework. *Journal of Computer Assisted Learning*, 37(6). <https://doi.org/10.1177/1071181321651253>



Human-Centered Learning Ecosystems: Reimagining Water Education for Real Estate Professionals

Danielle Storey, Poorva Ketkar, Emma Noble, & Harry Cooper

Continuing Education

Human-centered Design

Instructional Design

learning engineering

scenario-based learning (SBL)

User Experience

In Arizona, a gap exists between state-sponsored education and professional real estate practice, particularly in how licensed agents and brokers navigate water-related complexities such as rights, supply, and regional regulation. This paper presents the REAL Water Arizona project, a learning engineering initiative developed by the Arizona Water Innovation Initiative (AWII) at Arizona State University in partnership with the Arizona Department of Real Estate (ADRE). This work frames REAL Water Arizona as a learning engineering initiative that applies iterative, data-informed design cycles to transform compliance-driven continuing education into a scalable, human-centered learning system. Grounded in community engagement, user research, and applied instructional and interface design, the project seeks to enhance water literacy among Arizona's 83,000+ real estate professionals. Through evidence-based curriculum development and an interactive digital learning platform, researchers and designers advance the integration of

learning science, human factors, and public-sector collaboration to improve continuing education and professional decision-making in the context of Arizona real estate.

Introduction

The REAL Water Arizona project addresses a critical educational gap in the state's real estate sector. Many licensed real estate professionals, often the first point of contact for home buyers and sellers, lack confidence in discussing Arizona's complex water policies and regional supply realities. The Arizona Department of Real Estate (ADRE), which regulates real estate licensing and education, has recognized this gap by mandating a one-hour water course for all licensees effective January 1, 2025. In collaboration with ADRE, the AWII team seeks to improve statewide water literacy among Arizona's 83,000+ licensed agents and brokers through evidence-based curriculum design and digital learning innovation.

Arizona's water governance is regionally specific, with laws and management frameworks varying across boundaries like Active Management Areas (AMAs), Irrigation Non-Expansion Areas (INAs), and water unregulated zones. Given the complexity of Arizona's water systems, it is pedagogically and administratively unreasonable to expect individual CE instructors—many of whom are practicing agents to stay current on evolving hydrologic, legal, and policy developments. Applying a learning engineering approach (Goodell & Kolodner, 2022), the project converts this challenge into an opportunity to create a self-paced, expert-maintained learning system. Rather than treating curriculum development as a linear content-authoring task, REAL Water Arizona adopts a learning engineering mindset in which research, design, implementation, and evaluation are continuously aligned. We will deliver a curriculum co-developed by subject matter experts (SME) that integrates real-world relevance, quality user experience for licensees, learning science, and learning engineering principles to make water education both accessible and engaging.

Guided by the tenets of learning engineering interdisciplinary collaboration, technology as a tool to scale learning, design informed by learning science, and data instrumentation as a feedback loop (Goodell & Kolodner, 2022), the study investigates how water literacy can be improved through adaptive, research-informed instructional design driven by three questions:

1. What are the current water-related knowledge gaps among real estate professionals?
2. How can human-centered design improve engagement and comprehension of topics?
3. What data-driven methods optimize learning effectiveness for educators and learners?
4. Responding to higher education's call to bridge the "historic chasm between learning research and teaching practice (Thille, 2006), the ASU team serves as a learning engineering partner to ADRE. Through a tiered quality assessment and iterative design process, REAL Water Arizona brings together instructional designers, UX researchers, developers, hydrologists, policy experts, and ADRE staff to build a data-informed, contextually relevant learning platform for water literacy in real estate practice.
- 5.

Triangulation of Methods; Course observations, Surveys, Workshops, and Interviews

Our team used a mixed-methods approach to establish a holistic view of the landscape of water education for real estate professionals in both rural and urban areas of Arizona. A combination of four different research methods was designed to

ensure statewide reach, as well as qualitative and quantitative data encapsulating the lived experiences of real estate practitioners. Data was collected from professionals with active licenses who represent clients in real property transactions. Four team members observed 18 CE courses on water. Based on (Reinhart & Banister, 2018), a set of criteria was used to evaluate each class for its content coverage, materials/presentation quality, learner engagement, and depth of understanding. We conducted 14 semi-structured interviews to gauge the baseline knowledge, teaching practices, client concerns, past experiences, and recommendations for CE improvement. Interviewees were real estate professionals from agents from residential and commercial background, educators, and brokers who transact in a variety of different regulatory water locations.

To assess the effectiveness and perceived usefulness of current water courses among real estate professionals, we conducted nine 2-hour workshops across the state for CE credit. Workshops included interactive learning through small-group, scenario-based activities in which attendees developed client disclosure checklists related to water and real property. As evidenced by learning science studies, this deeper level of reasoning enhances learning gains (Craig et al., 2006). Drawing on Richter and Richter's (2024) assessment of professional development programs for teachers, workshop designs integrate cognitive activation, engaging prior knowledge through thought-provoking questions/reflections, and practical relevance, connecting course content directly to agent/broker workflow. We used a pre-workshop survey to collect participant demographics, areas of practice, perceived importance of water knowledge, confidence levels related to advising on water, client concerns about water, and water knowledge. This survey complements qualitative findings with quantitative ones.

Analysis Approach

Qualitative data, including the interviews and workshops, were thematically coded for patterns in common concerns, challenges, misconceptions, engagement strategies, and learner needs. Survey data were analyzed descriptively to quantify baseline knowledge and confidence levels. Notes from the course observation rubrics were transferred to an analysis framework examining the following areas: Instructor/Learner & Source Material Relations (how instructor and learner perceptions interface with water education content), Practical Relevance of Water to Real Estate Practice (how water education and real estate workflow are jointly represented), and Considerations for Student Needs (pedagogical approaches and supplemental resources). Insights were triangulated across methods to derive system-level themes, inform a taxonomy for the current state of educator approaches, highlight future content curation, and prioritize prototype features for the digital instruction platform.

Results

Data were collected from 83 licensed real estate professionals across Arizona. Rather than serving solely as descriptive outcomes, these findings function as design signals within a nested learning engineering cycle. The team:

1. Analyzed 18 ADRE-accredited CE courses to evaluate content accuracy, timeliness, and instructional design quality
2. Conducted 14 semi-structured interviews with educators and both residential and commercial practitioners to capture instructional challenges, misconceptions, and opportunities for improvement
3. Facilitated nine regional workshops across urban and rural regions of the state. Within these, 68 attendees participated in a structured survey helping to assess baseline knowledge, confidence, and perceived training needs

Findings revealed substantial knowledge gaps among both agents and educators in various locations, in topics including groundwater law, water supply designations and certificates, water rights, and water governance. Preliminary results are organized around key themes:

Regional Relevance and Contextual Learning

One of the most important findings was a strong regional variation in content needs. Workshops held in different parts of the state showed that water-related education must be localized based on regional governance, hydrology, and even the lived realities of each area (i.e. wells running dry). Different sessions had participants who spoke about certain topics and sparked more conversations specific to that area.

For instance, in urban/AMA regions, agents discussed rental subdivisions without 100-year supply requirements and Tribal water leasing/partnerships. In rural/INA regions, agents highlighted hauled water, seasonal wells, and a lack of client awareness. In southern Arizona, participants requested deeper coverage of stream adjudications, aquifer recharge, and groundwater management. Learners require localized content aligned with their areas. This finding directly informs both the instructional design framework and the organization of the web portal, ensuring that water education materials are differentiated by basin, water authority, and regional policy context.

Curriculum Consistency

Across 18 observed courses, the average total rubric score was 54.17 out of 75, indicating that most courses “met expectations” but showed a limited connection to practical workflow.

Table 1.
Average Rubric Scores for Observed Water Continuing Education Courses

Criteria	Average Scores	Highest Possible
ADRE Curriculum Topic Count	5.5	7.5
Course Content Coverage	12.39	20
Instructor’s Ability to Answer Questions	11.61	20
Source Credibility and Timeliness	14.39	20
Quality of Presentation Materials	12.61	20

The average inclusion rate was 73% across all ADRE-recommended topics. Foundational topics such as Groundwater and Identifying the 7 AMAs showed up in 100% of courses, while advanced or policy-related topics like Arizona Water Priority System (50%) and Proposed Commingling for Water Certificates (11%) were covered less frequently. Observation notes from workshops and courses showed that traditional lecture-heavy CE classes maintained low engagement, limited support, and topic inconsistencies. In contrast, our workshop design emphasized scenario-based learning, myth-or-fact games, and visual scaffolding (e.g., maps and infographics), producing higher participation during the sessions.

Baseline Knowledge and Misconceptions

Survey data and workshop discussions show measurable knowledge gaps and misconceptions about Arizona’s water systems. Among the 68 survey respondents, 81% were full-time real estate professionals, primarily agents and brokers, and 19% were educators. While most participants rated water knowledge as “highly important” to their practice, only 9% reported discussing water with clients daily, and 56% said they do so monthly.

The survey findings show that Arizona’s real estate professionals recognize water as an important issue influencing property value, client decisions, and long-term development, but have low confidence in communicating some topics, including groundwater management outside AMAs/INAs, water rights, and legal frameworks. To understand their baseline knowledge about certain topics, we asked them a few questions related to AMA/INA identification, water rights, and wells:

1. 38% (26) incorrectly identified Yuma and Kingman as AMAs, showing confusion about where AMA designations apply
2. 78% (36) incorrectly believed property water rights always transfer with land sales
3. 32% (15) incorrectly thought INA landowners can’t drill any new wells

Discussion

The REAL Water Arizona project illustrates how learning engineering can transform compliance-oriented continuing education into a dynamic, evidence-based learning ecosystem. By grounding design decisions in user research and iterative feedback, the project advances the learning engineering cycle of continuous data-informed improvement (Koedinger et al., 2021). Each workshop served as a small-scale design experiment, producing data on how real estate professionals learn complex, region-based water concepts.

Adaptive and Localized Learning

Water education in Arizona cannot succeed under a “one-size-fits-all” model. Regional variation in hydrology, governance, and lived experience requires modular and adaptive learning design, responsive to rapid changes at state water agencies, and also translated into the language of real estate. The platform’s regional modules, interactive maps, and a water themed chatbot for real-time inquiries address this challenge by aligning instruction with local realities and adapting content to user needs with generative AI (Huynh, Linh, & McNamara, 2025). Such contextualization supports situated learning, in which understanding emerges through authentic, practice-based contexts (Lave & Wenger, 1991).

Training the Trainers

Because ADRE-approved instructors act as the primary intermediaries between water experts and learners, “training the trainers” represents a crucial multiplier for learning outcomes. This approach aligns with human-centered learning engineering practices that position instructors as key stakeholders within the learning ecosystem rather than passive content deliverers (Thai et al., 2023). By equipping educators with modular lesson plans, video demonstrations, and ready-to-use digital materials, the project shifts the burden of technical expertise away from individual instructors and toward a sustainable, expert-maintained ecosystem. This design mirrors cognitive activation strategies that engage prior knowledge and promote reflection (Richter & Richter, 2024), while supporting conceptual change among learners (Posner et al., 1982).

Learning Engineering Implications and Future Work

REAL Water Arizona demonstrates how interdisciplinary collaboration can address complex, real-world learning challenges. The project’s methodology operationalizes human-centered research, instructional design, and data science to improve learning at scale (Goodell & Kolodner, 2022). Translating qualitative insights into an adaptive digital experience requires embedding feedback loops at every phase of development. At the beginning of this project, we established a feedback loop through mixed methods research to build a taxonomy of current educator practices. Extending this educator engagement into the design cycle and platform testing creates the opportunity to observe ways in which this taxonomy changes after a learning intervention. Thus aligning with a top priority of Learning Engineering, studying which practices in professional learning are effective and scalable (Baker, Boser, Snow, 2022). To ensure that the curriculum remains technically current while pedagogically relevant, our team will engage with subject matter experts and current CE educators to co-author learning

modules. In alignment with the Society for Learning Analytics Research framework (Giannakos et al, 2022), data instrumentation will be theoretically grounded, measuring constructs that matter to learning, not merely tracking platform activity.

Conclusion

The REAL Water Arizona project demonstrates how learning engineering can strengthen professional education by aligning curriculum design with real user needs. Mixed methods research identified knowledge gaps, current pedagogical practices and inconsistencies, and ways to increase content relevance through region-specific curricula. Future work will focus on piloting the online learning platform, consisting of interactive videos, micro learning modules, and shareable client resources. By integrating regional context, data-driven design, and interdisciplinary collaboration, the project introduces an innovative approach to continuing education that aligns Arizona's water literacy goals with the realities of professional practice. This learning engineering process lays the groundwork for a future model of continuing education that connects regulatory requirements with human-centered, evidence-based learning.

Acknowledgments

Amber Hedquist, Vriddhi Shah, and Vandhana Balaji for their support, collaboration, and contributions throughout this research.

Support provided by ASU's Arizona Water Innovation Initiative and the Impact Water - Arizona program, courtesy of the State of Arizona and the Virginia G. Piper Charitable Trust.

References

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*. <https://doi.org/10.1037/tmb0000058>
- Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). The Deep-Level-Reasoning-Question Effect: The Role of Dialogue and Deep-Level-Reasoning Questions During Vicarious Learning. *Cognition and Instruction*, 24(4), 565–591. https://doi.org/10.1207/s1532690xci2404_4
- Giannakos, M., Spikol, D., Di Mitri, D., Sharma, K., Ochoa, X., & Hammad, R. (2022). *The Multimodal Learning Analytics Handbook* (K. Sharma, X. Ochoa, D. Di Mitri, M. Giannakos, D. Spikol, & R. Hammad, Eds.; 1st ed.). Springer International Publishing AG. <https://doi.org/10.1007/978-3-031-08076-0>
- Goodell, J., & Kolodner, J. (2022). *Learning Engineering Toolkit: Evidence-Based Practices from the Learning Sciences, Instructional Design, and Beyond* (J. Goodell & J. Kolodner, Eds.; 1st ed.). Routledge. <https://doi.org/10.4324/9781003276579>
- Huynh, Linh, and Danielle S McNamara. "GenAI-Powered Text Personalization: Natural Language Processing Validation of Adaptation Capabilities." *Applied Sciences* 15, no. 12 (2025). <https://doi.org/10.3390/app15126791>
- Koedinger, K. R., Booth, J. L., & Klahr, D. (2021). The learning engineering approach: Building data-informed learning systems. *Educational Psychologist*, 56(3), 177–191.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

- Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66(2), 211–227.
- Reinhart, R. V., & Banister, S. (2018). Developing and Implementing Instrumentation for Digital High School Curricula: A Regional Study of a Rubric for Instructional Quality. *Athens Journal of Education*, 5(4), 361-374.
<https://doi.org/10.30958/aje.5-4-2>
- Richter, E., & Richter, D. (2024). Measuring the quality of teacher professional development: A large-scale validation study of an 18-item instrument for daily use. *Studies in Educational Evaluation*, 81, Article 101357.
<https://doi.org/10.1016/j.stueduc.2024.101357>
- Thai, K. P., Craig, S. D., Goodell, J., Lis., J., Schoenherr, J. R., & Kolodner J. (2023). Learning Engineering is Human-Centered. In J Goodell (Ed.), *The Learning Engineering Toolkit* (pp. 83-124). Routledge.
- Thille, Candace M. (2016). *Bridging Learning Research and Teaching Practice for the Public Good: The Learning Engineer*. New York, NY: TIAA Institute.



AI Systems Informed by Learning Engineering

A Predictive Learning Engineering Framework for Modeling Active Learning
Adaptive Multi-Modal Deepfake Detection for Safer Learning Environments
Agentic PAL: Designing Human-Empowered AI Partnerships for Early Childhood Mathematics Learning
An Application of Design-Based Implementation Research to Develop a Framework to Support a Community of AI-experience Creators
Automated Run-on Sentence Detection and Correction for Educational Writing
Automatic Identification and Evaluation of Revisions in Student Writing Using Large Language Models
Automatically Generating Interactive Learning Experiences with an LLM-Driven Agentic Pipeline
Bridging Human Intelligence Augmentation (IA) and Classroom Practices via GenAI in Learning Engineering
Charm-bots: The Impact of A.I.'s Sycophancy Language on User Trust
Coaching, Not Autocomplete: Early Evidence from ConnectInk's AI-Supported Personal Narrative Pilot
Comparing Epistemic Emotions and User Experience Across Two AI Instructional Designs in Biology Learning
Currents of Inquiry: Insights From Two Years of Real-World AI-Learner Water Conversations
Design and Pilot Evaluation of a Gamified Narrative Chatbot for STEM Education
Designing for Student Engagement with AI in Courseware: Lessons from Iterative Improvements to DOT

in REAL CHEM
EdLight Research Portal: An Expert-Annotated Repository of Handwritten Math Student Work
Epistemic Cognition and Uncertainty Navigation with a Domain-Specific AI Chatbot in STEM Education
From Course Concept to Lecture Video: An AI-Powered System for Automated MOOC Development
From Measurement to Action: A Learning Engineering Approach to AI-Powered Assessment for Human Power Skills Development
Implementing Concept Instruction via MCP Server
L2-French Learners and Generative AI (GenAI): Challenges, Needs, and Design Guidelines
Learning-By-Explaining with Generative AI: A Pilot Implementation in Introductory Biology
LLM Safety in an Educational Context: A holistic approach
Media Mentor AI: How a SCAMPER-guided AI assistant is helping reimagine media literacy learning
MIRANDA: Real-Time Learning Analytics for Authentic Embedded Assessment
Multiple-Document Comprehension in High School Science: A Learning-Engineering Pilot Study
NLP Validation of Prompt Strategies for Theory-Aligned LLM-Generated Personalization
Open Repository for AI Models as Learning Engineering Components
Optimizing Language-Focused Writing Feedback from Large Language Models through Prompt Engineering
Quality Assessment Through Learning Engineering: An Evaluation Rubric of LLM-Generated Multiple-Choice Questions
Reasoning LLMs are Competent Courseware Reviewers
ReQUESTA: A Hybrid Agentic Framework for Generating Cognitively Diverse Multiple-Choice Questions
Social and Emotional Dimensions of Generative AI Use
Socio-Emotional Learning in AI K-12 Guidance and Policy Documents: A Gap Analysis
The Difficult Conversations Bot: Findings on Fostering Empathy and Reflective Communication Among Faculty and Staff
The Writing Analytics Tool: A Learning Engineering Approach to Designing AI-Supported Writing Instruction
Towards Automated Detection of Struggling Student Programmers

User experience design of AI-assisted human-technology ecosystem for writing assessment

Using AI to Bridge Technology Gaps in Higher Education



A Predictive Learning Engineering Framework for Modeling Active Learning

Michelle Banawan, Tracy Arner, & Danielle S. McNamara

Active Learning

learning engineering

predictive analytics

Active Learning has been consistently linked to improved learning outcomes, persistence, and equity, yet it remains difficult to operationalize in digital and large-scale learning environments. While theory emphasizes internal cognitive processes such as sense-making, reflection, and strategic adaptation, learning-at-scale systems primarily capture observable behaviors. This disconnect has led to analytic approaches that equate engagement with behavioral presence, limiting the ability of learning analytics to represent meaningful learning activity. We present a predictive learning engineering framework that models Active Learning as an emergent, theory-aligned construct inferred from large-scale learner interaction data. Grounded in experiential, constructivist, sociocultural, and metacognitive theory, the framework conceptualizes Active Learning as patterns of generative, evaluative, and adaptive engagement. Using a human-centered, heuristics-guided construct specification process, predictive modeling, and iterative validation, the framework supports actionable and scalable active learning analytics in authentic learning-at-scale environments.

Introduction

Active Learning has been consistently linked to improved learning outcomes, persistence, and equity across instructional contexts. Despite broad consensus on its importance, Active Learning remains challenging to operationalize in digital and large-scale learning environments. This disconnect has led to analytic approaches that equate engagement with behavioral presence rather than meaningful learning activity. As a result, many dashboards and early warning systems struggle to distinguish productive struggle from disengagement, or strategic persistence from superficial activity. Learning engineering provides an opportunity to address this gap by systematically integrating theory, data, and iterative validation to produce actionable representations of learning processes.

Theoretical and Pedagogical Foundations

Active Learning is conceptualized as an instructional and cognitive process in which learners engage through purposeful activity, reflection, and feedback to construct and apply knowledge. This view traces to Dewey's principle of learning through reflective experience rather than passive reception (Bot et al., 2005). Piaget's theory of cognitive development situates learning as the active construction of knowledge (Barrouillet, 2015), while Vygotsky's sociocultural theory emphasizes learning through social interaction and scaffolding within the Zone of Proximal Development (McLeod, 2012). Kolb's experiential learning cycle

further articulates learning as an iterative process of experience, reflection, conceptualization, and experimentation. Within higher-education contexts, these theoretical commitments translate into pedagogical practices that intentionally engage learners in cognitive and social activity.

Predictive Learning Engineering Framework

Human-Centered Learning Engineering Orientation

The Predictive Learning Engineering Framework is grounded in the view that learning engineering is fundamentally human-centered, emphasizing deep understanding of learners, contexts, and systems rather than purely technical optimization. Consistent with Thai et al. (2022), learning engineering is conceptualized as an integrative discipline that draws on user-centered design, learning experience design, human–systems integration, design thinking, and universal design for learning to ensure that educational innovations meaningfully serve diverse learners. From this perspective, predictive analytics are not treated as abstract performance estimators, but as learning engineering instruments embedded within a broader instructional ecosystem. The framework begins with understanding the instructional challenge, the learner populations being served, and the contexts in which learning occurs. Empathy with learners, including recognition of their prior knowledge, strategies, needs, and constraints, guides construct specification and indicator design. This human-centered orientation helps ensure that operationalizations of Active Learning are theoretically anchored in experiential, constructivist, sociocultural, and metacognitive perspectives, reflect authentic learner experience rather than narrow behavioral proxies, and remain grounded in observable behavioral and cognitive processes that can be quantified at scale using multimodal educational data.

Construct Specification, Modeling, and Predictive Validation

To operationalize this structure at scale, we introduce the Learning Engineering Institute’s Predictive Analytics Framework, which models Active Learning as it emerges from authentic Learning-at-Scale (L@S) data through three interrelated stages:

- (1) Construct specification and indicator design. Grounded in learning theory and human-centered learning engineering principles, this stage identifies the core cognitive and behavioral processes underlying Active Learning and maps each construct onto feasible indicators supported by ASU’s L@S data environment. Because learning-at-scale data are inherently noisy and context-dependent, analytic heuristics guide indicator selection by prioritizing purposeful learner action over passive exposure, emphasizing temporal patterns rather than isolated events, ensuring interpretability for instructors and designers, and accounting for learner variability to avoid deficit-oriented assumptions.
- (2) Data-driven modeling of Active Learning signals. Using L@S datasets, we apply predictive and behavioral modeling techniques appropriate to the structure and granularity of available data to represent how students engage with courses at scale. Models leverage interaction patterns, submission behaviors, attempt trajectories, and participation indicators to capture meaningful variation in student learning journeys through composite and temporally sensitive signals.
- (3) Predictive validation. Modeled indicators are evaluated for coherence, predictive utility, and generalizability across courses and learner subgroups. Treating prediction as a form of empirical validation, this stage assesses whether emergent behavioral patterns correspond to meaningful learning processes and outcomes beyond what is captured by basic engagement metrics.

This framework exemplifies a learning engineering approach (Schatz et al., 2022). Construct specification functions as a design phase grounded in cognitive theory and heuristic guidance; data-driven modeling serves as iterative prototyping using real learner traces; and predictive validation provides continuous evidence for refinement and generalizability. Through this

engineering cycle, the framework produces actionable, theory-aligned indicators of Active Learning that can inform analytics, guide instruction, and support equitable, system-level decision-making.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305N210041 and R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the institute or the U.S. Department of Education.

References

- Barrouillet, P. (2015). Theories of cognitive development: From Piaget to today. *Developmental Review*, 38, 1-12.
- Bot, L., Gossiaux, P. B., Rauch, C. P., & Tabiou, S. (2005). 'Learning by doing': a teaching method for active learning in scientific graduate education. *European journal of engineering education*, 30(1), 105-119.
- McLeod, S. A. (2012, July). Zone of proximal development.
- Schatz, S., Thai, K. P., Craig, S. D., Schoenherr, J. R., Lis, J., & Kolodner, J. (2022). Human-centered design tools. In *Learning Engineering Toolkit* (pp. 279-301). Routledge.
- Thai, K. P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. (2022). Learning engineering is human-centered. In *Learning engineering toolkit* (pp. 83-123). Routledge.



Adaptive Multi-Modal Deepfake Detection for Safer Learning Environments

Syeda Samira Sama, Shubham Chakraborty, & Sami Mian

Adaptive Systems

Deepfake Detection

Educational Integrity

learning engineering

Multi-modal AI

Extended Abstract

Introduction

Generative AI now enables highly convincing deepfake audio and video, creating risks for education such as impersonated instructors, falsified assignments, and compromised proctoring. Although many detectors exceed 95% accuracy in laboratory tests, their performance often drops to 45–70% in real-world conditions due to dataset shift and rapidly evolving generative methods, limiting trust and adoption. From a Learning Engineering perspective, these failures represent a system-level breakdown in which learning technologies do not reliably support instructional integrity, trust, and equity. Our work asks: How can deepfake detection remain accurate, explainable, and fair enough for real educational deployment?

Approach or Design

We present a dynamic multi-modal ensemble framework that treats deepfake detection as an adaptive, data-informed service within learning platforms. The system integrates three specialized modes: Visual Sentinel for image artifacts, Acoustic Guardian for speech cues, and Temporal Oracle for video dynamics, each combining multiple state-of-the-art models. A gating network adaptively weights their contributions based on content characteristics and historical performance. Framed as a nested Learning Engineering cycle, the approach begins with identifying an authentic educational risk, implements multimodal AI solutions, evaluates performance in real learning contexts, and iteratively adapts the system based on empirical evidence. The pipeline updates itself as usage data and error patterns indicate where retraining, reweighting, or fairness adjustments are

needed, and human-centered design principles guide the use of explainable Grad-CAM visualizations to help instructors and administrators understand why content is flagged.

Findings or Insights

Across DF40’s 40 generation methods and several public benchmarks, individual modes achieve 96–98% accuracy, and the dynamic ensemble reaches 97.3% in laboratory settings while reducing false positives to approximately 1.3%. In real-world evaluations using Deepfake-Eval-2024, the system maintains 89.7% accuracy while baselines fall to 45–70%, reducing the laboratory-to-deployment performance gap from as high as 48% to 7.6%. Cross-modal verification identifies 73% of sophisticated deepfakes missed by single-modality detectors. Early fairness analysis shows 3–7% performance gaps across demographic groups, motivating further bias mitigation. We also received feedback from AAAI reviewers, which we plan to incorporate into the next stage of development. From a Learning Engineering standpoint, these findings illustrate how iterative evaluation across controlled and authentic contexts can reduce deployment risk and improve system trustworthiness in learning environments.

Table 1: Performance Metrics of Multi-Modal Ensemble Modes

Table 1 summarizes the performance metrics for each of the three specialized detection modes. By utilizing a dynamic weighted ensemble of transformer-based architectures, the system achieves high precision and AUC across all modalities. These high levels of accuracy provide a reliable foundation for maintaining educational integrity in digital learning environments where generative AI risks are prevalent.

Model Ensemble	Accuracy	Precision	Recall	F1-Score	AUC
Visual Sentinel (Image)	97.9%	98.2%	97.5%	97.8%	0.993
Acoustic Guardian (Audio)	96.2%	96.8%	95.7%	96.2%	0.984
Temporal Oracle (Video)	97.3%	97.8%	96.9%	97.3%	0.991

Implications

This work shows how adaptive ensemble architectures and cross-modal verification can turn deepfake detection into a continuously improving service suited to real educational needs. The framework can be integrated into learning management systems, virtual classrooms, and proctoring tools to verify the authenticity of student video submissions, provide explainable alerts rather than opaque classifications, and adapt as new generative models emerge. More broadly, the project reflects a learning-engineering cycle in which an educational risk is identified, multimodal AI methods are applied, real-world robustness and fairness are evaluated, and interpretable insights guide policy and instructional practice.

References

Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a nested learning engineering methodology to develop a team dynamic measurement framework for a virtual training

environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 conference proceedings: Solving for complexity at scale (pp. 115–132). <https://doi.org/10.59668/2109.21735>

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations. <https://arxiv.org/abs/2010.11929>

Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. Journal of Military Learning. <https://www.armyupress.army.mil/Journals/Journal-of-Military-Learning/Journal-of-Military-Learning-Archives/Conference-Edition-2023-Journal-of-Military-Learning/Engineering-at-a-Glance/>



Agentic PAL: Designing Human-Empowered AI Partnerships for Early Childhood Mathematics Learning

Anastasia Betts, Benjamin Ryon, Elida V. Laski, & Sunil Gunderia

Agentic AI

Early Childhood Mathematics

Family Engagement

Human-centered Design

learning engineering

Smart Learning Systems

The rapid expansion of AI-powered edtech has outpaced the field's capacity to ensure developmental appropriateness in early childhood mathematics. General-purpose large language models (LLMs) often generate activities that conflate concepts, skip developmental steps, or misalign with learning progressions, reflecting structural gaps in early childhood math expertise rather than limitations of AI itself. This paper introduces Agentic PAL, an AI-based system under development that uses a learning engineering approach to integrate learning science with a comprehensive, research-grounded early math knowledge infrastructure to support adult-child real-world math interactions. The theoretical foundations and emerging architecture behind Agentic PAL illustrate how thoughtfully designed AI-embedded edtech can strengthen adult expertise and learner outcomes in early childhood learning ecosystems.

Anastasia Betts¹, Benjamin Ryon², Elida V. Laski³, and Sunil Gunderia⁴

¹ Learnology Labs; anastasia@learnologylabs.org

² Learnology; ben@learnology.ai

³ Boston College; elida.laski@bc.edu

⁴ Mindset Copilot; sunil@mindsetcopilot.com

Abstract.

Keywords: Learning Engineering; Early Childhood Math; Agentic AI; Human-Centered Design; Smart Learning Systems; Family Engagement

Introduction

The rapid expansion of artificial intelligence (AI) in early childhood education has highlighted a growing tension between technological capability and the developmental nature of young children's learning (Hirsch-Pasek et al., 2015). Generic large language models (LLMs), now often embedded in "AI-powered" tools for learners, teachers, and families, frequently generate activities and explanations that misalign with early developmental learning progressions and fail to provide adults with the pedagogical supports needed to engage children in effective, developmentally grounded mathematics interactions (Betts,

2025). These challenges may not necessarily reflect the limitations of AI itself, but rather the absence of domain-specific knowledge infrastructures and pedagogical guardrails.

Agentic PAL (Personal Assistant for Learning) is an AI system that addresses this gap by integrating learning-science theory, learning engineering methods, and evidence from real-world use to empower adult-guided early math learning. Unlike traditional “AI tutors” that aim to instruct children directly on devices (Létourneau et al., 2025), Agentic PAL is designed to support (rather than replace) adults in helping young children learn through everyday real-world interactions. By functioning as a personalized learning concierge, Agentic PAL draws on a robust, continuously updated, research-grounded knowledge infrastructure to: (1) interpret informal adult observations and feedback captured during natural language conversations through a chat interface with an AI, (2) guide AI reasoning and system adaptivity, (3) support context-sensitive scaffolding, and (4) increase communication between home and school by sharing developmentally relevant, personalized data-driven insights for each learner.

Literature Review and Theoretical Grounding

Early mathematics learning emerges not from formal instruction alone, but through rich, frequent interactions with significant adults in everyday environments. Consider that children’s early exposure to mathematical talk, guided exploration, and hands-on math activity is strongly associated with later achievement (e.g., Levine et al., 2010; Gunderson & Levine, 2011). Moreover, high-quality adult–child interactions sharpen children’s attention to mathematical ideas (Daucourt et al., 2021). Effective systems aimed at supporting early mathematics should strive to strengthen the capacity and confidence of the adults who make these interactions possible, positioning adult expertise as a key leverage point for learning. Despite their critical role, many caregivers and early childhood educators report limited confidence in supporting young children’s mathematics learning (Ban et al., 2024). Math anxiety among parents is well-documented (Herts et al., 2019), and early childhood teachers often receive minimal preparation in early mathematics pedagogy or developmental learning trajectories (Clements & Sarama, 2020). These factors contribute to inconsistent or missed opportunities for mathematical learning in both home and school settings (Lu, Vasilyeva, & Laski, 2025).

Given this, the emergence of general-purpose large language models (LLMs) has fueled a rapid expansion of “AI-powered” tools for families and teachers. However, LLMs trained on broad internet data consume inconsistent and often inaccurate representations of early mathematics, including common confusions around foundational concepts (Betts, 2025). The resulting outputs include activities or explanations that are developmentally inappropriate or misaligned with learning progressions, and which provide limited guidance on how adults should scaffold, adapt, or interpret children’s thinking. These limitations highlight a critical learning engineering gap: early childhood mathematics requires AI systems that integrate learning-science principles, structured knowledge models, and explicit support for adult mediation in the real-world.

The learning sciences emphasize that young children progress through conceptual pathways in which later ideas depend on effective mastery of earlier ones (Bloom, 1984; Doignon & Falmagne, 1985), and that efficient progress along these pathways is supported through guided participation with a more knowledgeable other (Vygotsky, 1986; Zaretskii, 2009). Put simply, children learn most efficiently when adults offer timely scaffolds that extend their thinking just beyond what they can do independently, keeping learners in their Zone of Proximal Development (ZPD). This requires adults to recognize not only what the child currently understands, but what they are most ready to learn next (ZPD), and how to adjust support in the moment. In practice, reliably supporting children in this way places substantial cognitive and pedagogical demands on adults—demands that are rarely systematically or effectively supported in early childhood contexts.

These issues are further amplified by a “triple expertise” gap in early childhood mathematics (Betts, 2025): (1) a practitioner gap, in which many early childhood educators receive limited training in mathematical development (Clements & Sarama, 2020; Laski et al., 2013); (2) a perception gap, in which adults underestimate the complexity and importance of early math (Bittner & Bull, 2022); and (3) an infrastructure gap, in which the field lacks widely available, fine-grained knowledge models that map the developmental dependencies and common misconceptions (Clements & Sarama, 2024). Addressing these gaps

requires extensive research-based knowledge infrastructures (Doignon & Falmagne, 1985) that include systematic representations of concepts, learning trajectories, prerequisite relationships, and pedagogical strategies that encode what is known from decades of early mathematics research (Clements & Sarama, 2020). When operationalized through a learning engineering approach, such infrastructures prevent generative AIs from perpetuating the limitations of their training data, and enable them to effectively and accurately reason about learner readiness, interpret informal evidence from adult-child interaction feedback, and generate developmentally aligned personalized guidance for each learner.

A Learning Engineering Approach to Agentic PAL

Agentic PAL's design approach reflects the core commitments of learning engineering: integrating the learning sciences, using human-centered design methodologies, and using data-informed decision making to iteratively refine PAL's learning system (Goodell, 2023). The evolution of Agentic PAL is grounded in a nested learning engineering cycle (Craig et al., 2025) in which caregiver–AI interactions generate learning evidence that enables PAL to update individual children's knowledge models, while aggregate patterns of use inform ongoing system design and refinement.

PAL's first iteration was intentionally low-fi: a lightweight, text-message–based system that delivered research-aligned early mathematics activities to families, paired with simple feedback prompts for parents to report how the activity went. The system was built on a comprehensive, granular knowledge model containing hundreds of early math learning objectives organized by domain and developmental progressions that mapped precursor and successor relationships (Clements & Sarama, 2020). Parents received targeted activities matched to their child's "readiness" to learn key mathematical concepts and skills, along with brief explanations of why each activity mattered and how to best support their child's mathematical thinking (i.e., "Learning Science for Parents").

A small exploratory pilot study ($n = 19$) conducted during the Spring of 2025 showed that children whose families used PAL had greater gains on the TEMA-3 (Ginsburg & Baroody, 2003), with an average increase of 5.6 points compared to 1.9 points in a business-as-usual control group ($d > 1.0$). All PAL children showed improvement (100% vs. 57% in the control group), and 82% progressed from below to above grade level, with nearly all reaching kindergarten readiness and beyond by the posttest. Moreover, parent engagement was strongly associated with child learning gains ($r = .624$, $p < .01$; Betts, Ryon, & Laski, 2026). Building on this directional evidence, PAL is presently in use with ~500 students (from preschool through first grade), their families, and teachers across multiple sites in the United States. This larger-scale deployment is generating a rich stream of data and continues to affirm that adults benefit from regular, targeted support and that children's math learning increases when families engage. However, this broader real-world deployment has highlighted the limitations of a one-directional, prescriptive interface.

As the system has scaled, several new insights have emerged. Many families receive nudges at moments when they cannot engage, and although they may intend to return later, daily demands often intervene. Others complete activities but do not return to report how they went, leaving gaps in the evidence PAL needs to effectively adapt to a child's learning needs. Families also report frequent contextual mismatches; for example, an activity designed for the kitchen may arrive while they are at the grocery store, at the park, or waiting in the car, making it difficult to implement. In addition, many parents and early childhood teachers experience math anxiety or limited confidence in their own mathematical understanding (Herts et al., 2019; Lu, Vasilyeva, & Laski, 2025), which can make even accessible, informal activities feel intimidating.

Taken together, these realities highlight a fundamental limitation of traditional, structured input channels for capturing early learning in everyday contexts, and point to the kinds of challenges that thoughtfully designed AI systems are uniquely positioned to address. Systems intended to strengthen adult–child interactions must adapt not only to children's learning trajectories, but also to the affordances, constraints, and emotional landscapes of the adults who support them. Given this, a set of design requirements were identified based on close observation of how families and teachers interacted with PAL in real contexts (Table 1).

Table 1: Data-driven design requirements for Agentic PAL

Design Requirement	Description
R1: Flexible Timing	Adults must be able to share insights whenever they have capacity, not only when a prompt arrives.
R2: Natural Language Input	Parents and teachers need to describe interactions in their own words (in their own language), without structured reporting.
R3: Context Adaptation	Activities should adapt to real contexts (e.g., grocery store, car, park), not assume fixed settings.
R4: Interpretable Informal Evidence	The system must extract reliable learning signals from messy, everyday descriptions.
R5: Reduce Math Anxiety	Interactions should feel supportive and non-evaluative, building adult confidence.

Future Directions and Planned Research on Agentic PAL

Agentic PAL is currently moving through multiple phases of iterative design focused on developing a more dynamic and responsive solution, including: refining the conversational interface, strengthening natural-language evidence extraction, and operationalizing probabilistic inference mechanisms that map adult input onto the child’s individualized knowledge graph in PAL’s knowledge infrastructure. The next phase of work will compare the efficacy of Agentic PAL with the previous PAL (1.0) system. Across conditions, activities will remain aligned to the same underlying knowledge model, enabling the effects of the conversational, agentic layer on learning and family engagement to be examined. Data will include parent and teacher interactions with the system, process-level indicators of adult–child mathematical engagement, children’s learning trajectories within the knowledge model, and standardized assessment outcomes. This research will address several key questions:

1. Does Agentic PAL support more frequent or higher-quality adult–child mathematical interactions in everyday contexts?
2. Does natural-language evidence improve the accuracy and stability of the child’s evolving knowledge graph?
3. Do children whose parents use Agentic PAL demonstrate accelerated progress along early mathematics learning pathways compared to those supported by PAL 1.0?

Findings from this work will inform the iterative refinement of Agentic PAL going forward and contribute insights for others designing AI systems that empower humans, are grounded in the learning sciences, and who use learning engineering approaches to support learning across dynamic real-world contexts.

Conclusion

Rapidly evolving AI-powered tools in early childhood edtech underscore the need for a stronger, more explicit commitment to learning engineering approaches. Designing systems for young learners requires more than technical sophistication or generative flexibility; it demands grounding in the learning sciences and intentional design for how learning actually unfolds in everyday adult–child interactions. Effective systems must integrate theory, evidence, and iterative design in ways that strengthen the adults who mediate and guide young children’s early learning. Agentic PAL illustrates how these learning engineering commitments can be operationalized in practice. By combining a research-based knowledge infrastructure with a conversational, knowledge-guided interface, Agentic PAL is designed to help adults recognize what a child is ready to learn, how to scaffold effectively, and when and how to introduce appropriate levels of support and challenge. Importantly, Agentic PAL builds directly on evidence and insights generated through earlier implementations, demonstrating how learning engineering cycles connect field data, theory refinement, and system redesign. In this way, Agentic PAL serves not as a

singular solution, but as an example of how learning engineering approaches can help us move from insight to implementation, in direct alignment with the LERN Convening theme.

References

- Ban, J., Msall, C., Douglas, A. A., Rittle-Johnson, B., & Laski, E. V. (2024). Knowing what they know: Preschool teachers' knowledge of math skills and its relation to instruction. *Journal of Experimental Child Psychology*, 246, 105996.
- Betts, A. (2025, July 23). The AI mirror: How GenAI reflects and amplifies gaps in early math. The Learning Agency.
- Betts, A., Ryon, B., & Laski, E. (2026, April). "Smart" nudges, real gains: Family engagement for kindergarten math readiness. Paper to be presented at the Annual Meeting of the American Educational Research Association (AERA), Los Angeles, CA.
- Bittner, K., & Bull, R. (2022). Understanding educator attitudes towards and perceptions of mathematics in early childhood. *Journal of Early Childhood Research*, 20(3), 341-356.
- Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6), 4-16.
- Clements, D. H., & Sarama, J. (2024). Systematic review of learning trajectories in early mathematics. *ZDM – Mathematics Education*. Advance online publication.
- Clements, D. H., & Sarama, J. (2020). *Learning and teaching early math: The learning trajectories approach*. Routledge.
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale* (pp. 115-132).
- Daucourt, M. C., Schatschneider, C., & Hart, S. A. (2021). The home math environment and math achievement: A meta-analysis. *Journal of Experimental Child Psychology*, 205, 105073.
- Doignon, J. P., & Falmagne, J. C. (1985). Spaces for the assessment of knowledge. *International journal of man-machine studies*, 23(2), 175-196.
- Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability* (3rd ed.). PRO-ED.
- Goodell, J. (2023). Chapter introduction: What is learning engineering? In J. Goodell & J. Kolodner (Eds.), *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond* (pp. 5–13). Taylor & Francis.
- Gunderson, E. A., & Levine, S. C. (2011). Some types of parent number talk count more than others: Relations between parents' input and children's cardinal-number knowledge. *Developmental Science*, 14(5), 1021–1032.
- Guss, S. S., Clements, D. H., & Sarama, J. (2022). High-quality early math: Learning and teaching with trajectories and technologies. In A. Betts & K.-P. Thai (Eds.), *Handbook of Research on Innovative Approaches to Early Childhood Development and School Readiness* (pp. 349-373). IGI Global.

- Herts, J. B., Beilock, S. L., & Levine, S. C. (2019). The role of parents' and teachers' math anxiety in children's math learning and attitudes. In *Mathematics anxiety* (pp. 190-210). Routledge.
- Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015). Putting education in "educational" apps: Lessons from the science of learning. *Psychological science in the public interest*, 16(1), 3-34.
- Laski, E. V., Reeves, T. D., Ganley, C. M., & Mitchell, R. (2013). Mathematics teacher educators' perceptions and use of cognitive research. *Mind, Brain, and Education*, 7(1), 63-74.
- Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning*, 10(1), 29.
- Levine, S. C., Suriyakham, L. W., Rowe, M. L., Huttenlocher, J., & Gunderson, E. A. (2010). What counts in the development of young children's number knowledge? *Developmental Psychology*, 46(5), 1309–1319.
- Lu, L., Vasilyeva, M., & Laski, E. V. (2025). Home math environment as a mediator of socioeconomic differences in early math skills: A study of Chinese families from disparate backgrounds. *Developmental Psychology*.
- Thai, K. P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. L. (2023). Learning engineering is human-centered. In J. Goodell (Ed.), *The learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond* (pp. 83–124). Routledge.
- Vygotsky, L. S. (1986). *Thought and language*. MIT press.
- Zaretskii, V. K. (2009). The zone of proximal development: What Vygotsky did not have time to write. *Journal of Russian & East European Psychology*, 47(6), 70-93.



An Application of Design-Based Implementation Research to Develop a Framework to Support a Community of AI-experience Creators

Janice Mak, Siyuan Li, & Katerina Christhlf

Extended Abstract

Generative artificial intelligence (GenAI) has rapidly shifted the ways in which we communicate, interact, teach, study, and learn. The shift from being a consumer to a creator of GenAI experiences, however, is a tremendous transformation, whether teachers creating a bot for students to engage with to learn German (Reves & Bahtchevanova, 2025) or students building a customized bot to engage with their course content (Pearce et al., 2023). Our work with our institution's CreateAI Builder, a platform that allows faculty, staff and students to design and deploy customized bots has transformed the educational landscape and unleashed innovation in ways never previously imaginable. This led us to ask: What supports do creators need to create ethical, learner-centered AI-X tools with pedagogical value? How can we build a framework and accompanying tools to provide these supports? In this paper, we describe our design-based iterative research study and the participatory, co-design approach to develop and iteratively refine an AI-X Framework that integrates Principled Innovation (PI; Arizona State University, n.d.) and learning engineering (LE; Goodell & Kolodner, 2023) into three interconnected tools that support different stages of AI-X development. We gathered input from a set of design team working sessions regarding key ethical, pedagogical, and design considerations regarding AI-X. We then conducted in-person, synchronous, and asynchronous academic-year feedback sessions that drew on representation from the entire institution across colleges and departments to iteratively refine the AI-X Framework. The project integrates two anchoring bodies of work: PI (Arizona State University, n.d.) and LE (Goodell & Kolodner, 2023).

Acknowledgments

The research described herein has been sponsored by the Kern Family Foundation. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the Kern Family Foundation, and no official endorsement should be inferred.

References

Goodell, J., & Kolodner, J. (2023). *The Learning Engineering Toolkit: Evidence-based practices from the Learning Sciences, Instructional Design, and beyond* (1st ed.). Routledge.

Pearce, K., Alghowinem, S., & Breazeal, C. (2023). Build-a-bot: teaching conversational AI using a transformer-based intent recognition and question answering architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 13, pp. 16025-16032).

Reves, C., & Bahtchevanova, M. (2025, September 23). *Designing with Purpose: Pedagogical*

Reflections for the Development of AI Speaking

Tools. The FLTMAG. <https://fltmag.com/pedagogical-reflections-ai-speaking/>



Automated Run-on Sentence Detection and Correction for Educational Writing

Shubham Chakraborty, Yu Tian, Michelle Banawan, Andrew Potter, Linh Huynh, Yoshita Yajjapurapu, & Danielle S. McNamara

Grammatical Error Correction

learning engineering

natural language processing

Writing Analytics

Run-on sentences, including fused sentences, comma splices, and conjunctive adverb misuse, pose a persistent challenge in student writing, undermining both human evaluation and automated analyses in learning environments. Despite their instructional importance, run-ons are underrepresented in major grammatical error correction (GEC) benchmarks. We present a two-stage NLP pipeline for run-on detection and minimal-change correction, designed within the Learning Engineering Framework to improve writing feedback while preserving student voice. Early annotation of 251 student sentences identified 29 potential run-ons, informing our pipeline design and human validation workflows.

Introduction

Run-on sentences represent a high-impact yet underexplored error type in educational writing. Within the Learning Engineering Framework (Goodell & Kolodner, 2023), this work creates data-informed systems that improve writing feedback while preserving student voice.

Run-on sentences occur in three primary subtypes: Comma Splice: Two independent clauses joined only by a comma without a coordinating conjunction (e.g., "I studied all night, I still failed the exam"). Fused Sentence: Two independent clauses with no punctuation or conjunction between them (e.g., "She finished her essay she submitted it online"). Conjunctive Adverb Misuse: A comma incorrectly placed before a conjunctive adverb connecting two independent clauses (e.g., "The results were promising, however more research is needed").

Despite their instructional importance, run-ons are underrepresented in major GEC benchmarks such as BEA-2019 and CoNLL-2014 (Bryant et al., 2019), which emphasize broad error categories over clause-boundary errors common in student writing. Consequently, AI writing tools inadequately address run-ons, and no systematic framework exists for modeling this error type with minimal, voice-preserving correction.

Approach

Two-Stage Pipeline Architecture

We present a two-stage pipeline guided by iterative prototyping, structured data instrumentation, and linguistic theory.

Stage 1: Run-on Detection. The first stage focuses on classifying sentences as run-ons and identifying subtype categories. We employ transformer encoders (BERT, RoBERTa, DeBERTa) fine-tuned on annotated student writing samples. The detection model learns to recognize clause boundaries, punctuation patterns, and syntactic structures indicative of run-on errors. Subtype classification enables targeted feedback that helps students understand the specific nature of their error, moving beyond generic "grammar error" notifications toward pedagogically meaningful guidance.

Stage 2: Minimal-Change Correction. The second stage generates corrections using sequence-to-sequence models (T5, FLAN-T5, LLaMA-3) fine-tuned on paired error-correction examples. The key design principle is minimal-change correction: producing the smallest grammatical modification necessary to resolve boundary errors while preserving the rhetorical intent and stylistic choices of the student writer. Each run-on subtype allows multiple valid corrections, and the model selects the most contextually appropriate one while preserving minimal edit distance from the original text. This approach respects student voice and avoids over-correction that could alter meaning or discourage developing writers.

Data Instrumentation

Training data derives from Canvas submissions, student essays, and discussion posts collected through Arizona State University's Learning at Scale (L@S) research infrastructure. Initial annotation uses GPT-based few-shot prompting followed by human expert validation. Trained annotators review machine-generated annotations to confirm or reject run-on classifications, verify subtype labels, identify edge cases requiring adjudication, and evaluate correction quality for voice preservation. This human-in-the-loop workflow ensures annotation quality while enabling efficient processing of large text collections. Inter-annotator agreement is measured using Cohen's Kappa to ensure reliability, with disagreements resolved through expert adjudication to create gold-standard labels. Subtype balance is maintained through targeted sampling and data augmentation to address class imbalance inherent in naturalistic writing samples.

Early Insights and Preliminary Annotation

To support early design decisions for the detection and correction pipeline, we conducted an initial study using GPT-4-turbo with few-shot prompting to generate preliminary annotations for 251 student sentences.

Table 1.

Preliminary annotation results from initial study (n=251 sentences).

Category	Count
Total sentences annotated	251
Potential run-ons identified	29
Detection rate	11.6%

The model identified 29 potential run-on sentences distributed across the three subtypes. Example constructions identified include comma splices in argumentative essays where students connect related claims, fused sentences in narrative passages with rapid action sequences, and conjunctive adverb misuse with transitional words like "however" and "therefore." These annotations help surface challenging constructions, prototype subtype heuristics, and refine human-in-the-loop validation workflows. The relatively low detection rate (11.6%) aligns with expectations for student writing at the undergraduate level, though validated rates will emerge from ongoing human annotation efforts.

Implications for Learning Engineering

This project contributes to learning engineering through four areas: Data Instrumentation, by providing robust annotation protocols for an underrepresented clause-boundary error; Learner-Sensitive Design, by prioritizing minimal-change corrections that known preserve student voice; Educational NLP Infrastructure, by delivering validated resources and a modeling pipeline that support writing analytics in instructional contexts; and FERPA Compliance, by ensuring all processing occurs on secure university infrastructure without external API calls.

Conclusion and Next Steps

We have presented a two-stage pipeline for run-on sentence detection and correction in educational writing, designed within the Learning Engineering Framework. Our preliminary annotation study identified 29 potential run-ons in 251 student sentences, informing pipeline design and validation workflows. Next steps include: (1) completing human expert validation to establish ground truth labels, (2) fine-tuning transformer models on validated annotations, (3) evaluating detection precision and recall against human judgments, and (4) assessing correction quality through both automated metrics and human evaluation of voice preservation. This work addresses a significant gap in educational NLP by targeting an error type that is common in student writing yet underrepresented in existing GEC systems.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305N210041 and R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

- Bryant, C., Felice, M., Andersen, Ø. E., & Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 52–75). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-4406>
- Goodell, J., & Kolodner, J. (Eds.). (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Taylor & Francis.
- Omelianchuk, K., Atrasevych, V., Chernodub, A., & Skurzhanskyi, O. (2020). GECToR – Grammatical error correction: Tag, not rewrite. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 163–170). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.bea-1.16>



Automatic Identification and Evaluation of Revisions in Student Writing Using Large Language Models

Yu Tian, Parul Patil, Andrew Potter, Jessica Early, Steve Graham, Katerina Christhlf, Motahareh Darvishpour Ahandani, & Danielle S. McNamara

Large Language Models

Revision Evaluation

Extended Abstract

Revision provides a window into writers' metacognitive awareness, goal-directed decision making, and linguistic development (Hayes, 2012). Understanding how students revise is therefore critical for both writing research and instruction. However, identifying and evaluating revisions remains challenging: manual annotation is labor-intensive and inconsistent, while existing automated approaches rely largely on surface metrics (e.g., deletion/insertion counts, edit distance) (Tian & Cushing, 2025). These limitations prevent researchers and educators from capturing the semantic and functional nature of revisions, ultimately constraining the usefulness of revision analytics for scalable, instructionally meaningful feedback. Advancing automated methods for identifying and evaluating revisions is essential for generating evidence-based feedback and supporting personalized writing development at scale.

This project introduces an automatic workflow that leverages large language models (LLMs) to identify, categorize, and evaluate revisions between students' draft and revised essays (See Appendix A for an overview). The workflow includes four stages: (1) detecting text changes between draft:revision pairs using rule-based approaches, (2) prompting an LLM to explain each change in natural language, (3) using an LLM to categorize revision types (e.g., typographic, grammatical, or content-level), (4) querying an LLM to evaluate each revision as Good, Neutral, or Bad and provide explanation. Two trained human annotators independently code revisions using a custom rubric (Appendix B). The annotations serve as ground truth for evaluating the accuracy and reliability of the automated workflow.

Initial analyses indicate that, with carefully designed prompts, state-of-the-art LLMs can reliably identify revision types and evaluate their quality using the textual context. We anticipate that our workflow will achieve performance comparable to human annotation and demonstrate strong potential for integration into writing research and instructional tools.

By harnessing LLMs for automated text revision analysis and evaluation, this work advances the development of evidence-based, scalable systems for writing analytics and personalized feedback. The workflow embodies the principles of learning engineering by uniting theory, data, and iterative design to model and support complex learning processes. Anticipated applications include formative feedback systems that highlight meaningful revisions, instructor dashboards that visualize students’ writing growth, and adaptive learning environments that guide students toward higher-quality revisions.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

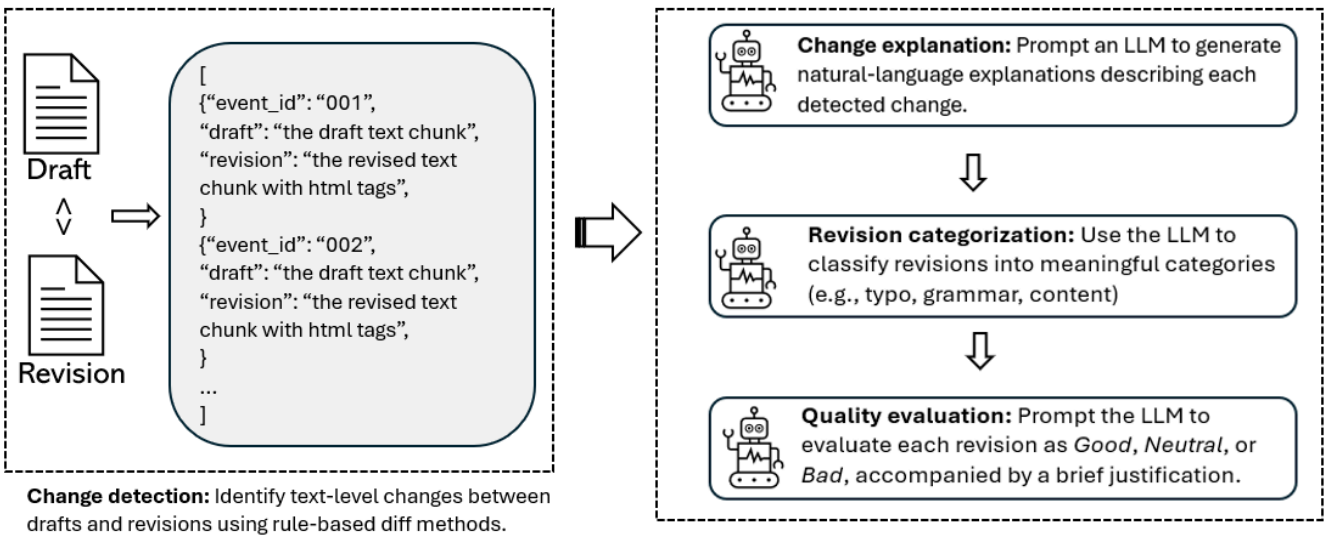
Hayes, J. R. (2012). Modeling and remodeling writing. *Written communication*, 29(3), 369-388.

Tian, Y., & Cushing, S. T. (2025). Exploring the application of keystroke logging techniques to research in second language (L2) writing. *Research Methods in Applied Linguistics*, 4(1), 100179.

Appendix A

Figure A1

Overview of the automated workflow for revision identification and evaluation.



Appendix B

Table B1.

Rubric for Revision Identification and Evaluation

Category	Definition	Decision Rule	Example
Typographic Revision (TYPO)	Changes that correct visual or mechanical surface errors not affecting meaning (e.g., spelling typos, missing/extra characters, accidental punctuation).	If the only difference is a spelling correction, punctuation fix, capitalization, or spacing and the semantic/pragmatic content is unchanged → TYPO.	"teh" → "the"; "its", → "it's"; extra double space removed
Grammatical & Conventions Correction (GRAMMAR)	Changes that fix grammar, morphology, tense/aspect, subject-verb agreement, word order that affect grammaticality but not content meaning substantially.	If the edit improves grammaticality or register (e.g., corrects run-on, reorders words to follow grammar) but does not add new content or change argumentative stance → GRAMMAR.	"He go to school" → "He goes to school"; "in the 1990s" → "during the 1990s"
Word Change (WORD)	Changes in word choice that adjust word-level meaning, specificity, or register without major structural reorganization. Includes synonym replacements, word-level clarifications, or idiom fixes.	If single-word or short phrase swaps change lexical precision, tone, or style, but not intended to change content or add new information → WORD	"increase" → "rise".
Sentence Restructuring (SENTENCE)	Revisions that reorganize clause/sentence structure without substantial addition/removal of content — e.g., splitting/combining sentences, passive→active, reorder of clauses for flow.	If the rewrite changes sentence boundaries or clause order to improve clarity/coherence but does not substantially change content facts → SENTENCE.	"Although it rained he went." → "He went, although it rained."
Content Revision (CONTENT)	Revisions that add, remove, or substantially change content. This is substantive: new ideas, elaboration, addition/removal of examples, thesis change.	If the edit alters meaning, adds new details, removes or changes the information → CONTENT.	"The study was inconclusive." → "The study failed to reach significance due to small sample size, suggesting further research." (addition/clarification).



Automatically Generating Interactive Learning Experiences with an LLM-Driven Agentic Pipeline

Felix Gröner, Anish Verma, & Jason Bronowitz

AI Pedagogy

Educational Technology

Instructional Design

Interactive Learning

Large Language Models

prompt engineering

Interactive and hands-on learning experiences encourage deeper engagement and produce better learning outcomes than passive consumption of media. We see an opportunity to improve learning by leveraging the ability of large language models (LLMs) to generate text and code for the automated production of interactive learning experiences. We present a chain of LLM agents that plan, generate, review, and refine a static interactive tutorial website for any topic. Their prompts were engineered to follow best prompting practices and employ pedagogic theory. This system is motivated by the effectiveness of interactive learning and the prospect of making education more accessible.

Introduction

Large language models (LLMs) are becoming increasingly integrated into educational contexts, particularly for the rapid generation of instructional content such as explanations, exercises, and summaries. However, much of AI-generated educational material remains pedagogically underdeveloped because it prioritizes fluency and surface-level coherence over meaningful learner engagement (Zawacki-Richter et al., 2019; Sharma et al., 2025). Many AI-supported learning resources reproduce passive formats that ultimately mirror textbooks or lecture notes and offer limited opportunities for learners to actively construct knowledge.

This limitation stands in contrast to decades of empirical research in the learning sciences which demonstrate that interactive and hands-on learning environments consistently outperform passive instructional formats. Prior work found that learners achieve stronger conceptual understanding, motivation, and retention when they are required to manipulate variables and explore systems dynamically (Koç & Kanadlı, 2025; Li et al., 2023).

In the context of the Learning Engineering Process cycle (Goodell & Kolodner, 2023), we take this challenge and report the results of the creation phase (designing and building a solution). We leverage the ability of LLMs to not only generate pedagogically-grounded instructional content but also automate the development of web pages that offer interactive and individually tailored learning experiences. Our iterative pipeline plans, evaluates, and refines the content quickly and cheaply, promising to make education more easily accessible.

Theoretical Foundations

In the spirit of Learning Engineering, we are not just developing technology but following the guidance of empirically backed science. Our design is anchored in complementary frameworks that articulate how learning occurs, how instruction should be structured, and how technology should be integrated into pedagogy.

The ICAP framework (Chi & Wylie, 2014) characterizes learning activities according to levels of cognitive engagement: passive, active, constructive, and interactive. It demonstrates that learning outcomes systematically improve as learners move toward higher levels of engagement. The distinction between mere activity and genuine interactivity is particularly relevant: Learning is most robust when learners are required to generate, test, and revise ideas through interaction.

Mayer's multimedia learning theory provides further constraints on instructional design by specifying how information should be presented to align with human cognitive architecture (Mayer, 2021; Clark & Mayer, 2016). Our design follows the principles such as coherence, signaling, and segmentation to reduce extraneous cognitive load and support meaningful processing.

The Technological Pedagogical Content Knowledge (TPACK) framework (Mishra & Koehler, 2006) situates these cognitive principles within a broader systems perspective, emphasizing that effective instruction emerges from the integration of content knowledge, pedagogical knowledge, and technological design.

System Design

We use LangGraph to chain together several LLM invocations, create iteration loops, and run processes in parallel. We use a combination of OpenAI's GPT-5 mini and Google's Gemini 3 Flash as they proved most practical because they are fast, cheap, and easily accessible. We found that other models did not produce results of sufficient quality. OpenAI's and Google's models were practically interchangeable for most agents with the exception of the initial website generation for which OpenAI's results were unsatisfactory.

In a first step, the user-given topic and additional info are provided to a Tutorial Planner to formulate learning goals, compose an outline for the tutorial, and suggest interactive elements.

This plan is then passed to the Website Generator which generates a first draft of the static web page. We decided to separate these two tasks as specialized prompts generally produce better outputs, even though today's LLMs have sufficiently large context windows for much longer prompts. While it is generally best practice to provide the LLM with examples, we found that a complete finished page amounted to more tokens than the rest of the prompt, threatening to overpower our instructions. Instead, we insert an empty code template that includes predefined styles to ensure stylistic consistency.

We then iterate over this draft in a loop where first, the website is checked for syntax errors by ESLint and for runtime errors by Playwright. The latter also takes a screenshot of the page. These are then fed into three parallel critics: a Code Reviewer, a Visual Critic, and Pedagogy Expert. We found that separating this review into three specialized critics drastically improved the ratio of identified issues. The LLMs are prompted to produce a report and also state whether critical issues need to be resolved before publishing the page. If any reviewers identify critical issues, the reports and the code are fed into a Code Fixer and the loop starts over.

Findings, Implications, and Limitations

Preliminary results suggest that the proposed pipeline is capable of producing functional and conceptually coherent interactive tutorials across a range of topics. This novel technology could produce low-cost, individually tailored teaching applications at the learner's convenience, making effective learning more accessible than ever.

Our development experience aligns with the common finding that narrowly specialized agents outperform general-purpose prompts, particularly when reviewing different aspects such as code quality, aesthetics, usability, and pedagogical value. We also observed meaningful variation across 5 OpenAI and 4 Gemini LLMs we tested, with some models excelling at web design and others performing better in visual critique.

We assume that there are many domains for which our interactive learning experiences are not the best method. Because our pipeline aims at increasing understanding, skills that need to be practiced are not suited for our tutorials. We also found that limitations of the browser caused issues when extensive calculations are required (e.g., Fourier transform). Another limitation of our approach is the qualitative testing and evaluation during the iterative prompt engineering process. The low number of test outputs usually exhibited a high within-model variance, making it difficult to objectively and robustly assess the performance of prompts.

Future work should move this project to the next phase of the Learning Engineering Process cycle by implementing and investigating such generated tutorials with real learners.

References

- Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243. <https://doi.org/10.1080/00461520.2014.965823>
- Clark, R. C., & Mayer, R. E. (2016). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning* (4th ed.). Wiley.
- Goodell, J., & Kolodner, J. (Eds.). (2022). *Learning Engineering Toolkit: Evidence-Based Practices from the Learning Sciences, Instructional Design, and Beyond*. Routledge. <https://doi.org/10.4324/9781003276579>

- Koç, A., & Kanadlı, S. (2025). Effect of interactive learning environments on learning outcomes in science education: A network meta-analysis. *Journal of Science Education and Technology*, 34, 681–703. <https://doi.org/10.1007/s10956-025-10015-7>
- Li, M., Ma, S., & Shi, Y. (2023). Examining the effectiveness of gamification as a tool promoting teaching and learning in educational settings: A meta-analysis. *Media Psychology*, 26(4), Article 1253549. <https://doi.org/10.1080/15213269.2023.1253549>
- Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). Cambridge University Press.
- Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017–1054.
- Sharma, S., Mittal, P., Kumar, M., & Bhardwaj, V. (2025). The role of large language models in personalized learning: A systematic review of educational impact. *Discover Sustainability*, 6, Article 243. <https://doi.org/10.1007/s43621-025-00243-7>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education*, 16, Article 39. <https://doi.org/10.1186/s41239-019-0171-0>



Bridging Human Intelligence Augmentation (IA) and Classroom Practices via GenAI in Learning Engineering

Li (Lee) Liang

Artificial Intelligence

Classroom Practices

Human IA (intelligence augmentation)

learning engineering

Personalized Learning

UX Design

This paper is written to share a grounded and practice-oriented understanding of Dr. Christopher Dede's keynote series 2022-2025 on human intelligence augmentation (IA) via learning engineering in response to his invitation for public feedback. The goal here is to share a practice-oriented understanding of the significant usefulness of this epistemological framework on bridging human IA, GenAI, and classroom practices via learning engineering. This paper synthesizes key epistemological themes from Dede's IA keynotes (2022–2025) and interprets their implications for classroom practice and learning engineering. This paper also discusses some major real-world opportunities, pitfalls and possible solutions to improve the educational ecosystems to implement IA in classrooms via learning engineering for the front-line educators, policy makers, researchers, learning engineers and AI developers, etc. Every reader is advised to create

their own version of user experience (UX) design for human IA to benefit their work.

Li (Lee) Liang

Lee Vision and EvalMate.AI; llcfy2022@gmail.com

Abstract.

Keywords: Intelligence Augmentation (IA), Artificial Intelligence (AI), Learning Engineering, UX Design, Personalized Learning, Classroom Practices.

Introduction

“Unless we are wise, LLM may become a type of digital duct-tape to hold together on obsolete industrial-era educational system.”

— Dr. Christopher Dede (2025, May 19)

I first discovered the profound value of Dr. Christopher Dede’s epistemology framework on human intelligence augmentation (IA) via learning engineering from his keynote speech titled “Intelligence Augmentation via Artificial Intelligence: IA rather than AI” in the third Empowering Learners in AI 2022 International Conference from 6-8th December 2022, just a week after ChatGPT’s official debut on Nov. 30th, 2022. This significant yet underestimated roadmap of human IA targets for educational user experience (UX) design and research via learning engineering to meet the new workforce demands for human survival and thriving in GenAI age. Dr. Dede is a pioneering researcher in learning engineering and AIED for decades. Having been working with AI for more than half a century, he has lived through nine hype cycles for AI development. His primary interest lies in developing innovative educational systems to tackle the 21st century’s opportunities and challenges. This paper aims to understand this question: how to bridge the gap in classroom practice, human IA and learning engineering through synthesizing Dede’s keynote speeches in his website.

Methods

To ensure the credibility and rigor of this paper, the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Page et al., 2021) were adopted. Six keynotes on IA via AI from 2022 to 2025 were collected from Dede’s website: <https://www.chrisdede.com/videos>. Content analysis was adopted for the data analysis in the explanatory synthesis of Dede’s keynotes and the relevant learning engineering framework.

Results

Three core pillars of epistemology framework on IA are synthesized to identify and bridge the gap in classroom practice, human IA and learning engineering from Dede’s keynotes and the relevant learning engineering framework as followed.

Epistemological Framework

Bridging the gap in epistemological foundations of IA

The learning engineering process starts with a significant challenge (Kessler et al., 2023; Thai et al., 2023). Human IA as the optimal goal of UX design for AIED has been largely underestimated in GenAI age as all eyes have been focusing on the thrilling race to advance the augmentation of AI since the debut of ChatGPT in 2022. What would be the consequence if we do not prioritize human IA? Inevitably, the more intelligent AI, the less proficient humans. Unfortunately, this paradox has already become a reality, evidenced by the 2024 Nation's Report Card that 70% of 8th graders were below proficient in reading, and 40% of 4th graders did not meet basic reading levels (National Center for Education Statistics, 2025; Ginsberg & Zhao, 2025). Here one cannot highlight enough the importance, urgency and necessity of human IA via AI. In fact, just a week after the ChatGPT debut in 2022, Dede (2022, December 6) immediately alerted us about this risk and swiftly pointed out a very timely but undervalued solution path to fix this paradox problem in his keynote speech in 2022.

A stitch in time saves nine. So, to bridge the gap in epistemological foundation of IA, what workforce skills for human IA should we aim at? Dede (2022, December 6) suggested two theoretical frameworks for this purpose: the future cognitive skills 2030 from NESTA (Bakhshi, et al., 2017), and the essence of socio-ethical skills 2030 from OECD (2017).

For IA in university education, Dede (2022, December 6) called attention to the five most sought-after competencies by employers within the entry-level workforce reported in the Georgetown University Center on Education and the Workforce (2020). Thus, a category of workforce skills for human IA can be concluded: problem solving and higher-order thinking, leadership, communication, teamwork, selling and promotion.

For the overwhelming public misinformation and concerns that AI may wipe out many human jobs, it was clarified that it is more likely AI will reshape the division of labor across most professions, necessitating a shift in human workforce development toward emphasizing uniquely human skills that AI is not good at (Dede, 2022, December 6). Dede (2022, 2023, 2024) compared the differences between persons and AI in an effort to maximize human goods and minimize human bads with complementary assistance from AI.

Furthermore, Dede (2023, August 7) offered a balanced and evidence-driven example of what LLMs can and cannot do in educational contexts for the implementation phase in learning engineering process. An ideal complementary human-AI collaboration at a new level of efficiency in GenAI age will be having AI for the calculative reckoning tasks, freeing humans for the decision making and uniquely human tasks such as human taste or vision that often involve higher-order thinking such as critical thinking and creativity. Dede (2023, August 7) also noted that a shift in what to assess in GenAI age is particularly necessary since AI does well on the psychometric high-stakes summative assessment, including even the descriptive writing and art.

Next, LLMs may also bring us unforeseen pitfalls that easily mislead us. Given the fact that LLMs are trained on the existing knowledge base from the past, they are often not able to provide learners new knowledge for the very challenging and novel future ahead. Equally, they are not preparing to offer learners creative answers but formulaic ones from the past. They are just preparing teachers to generate lots of AI-driven standardized psychometric quizzes, predictable slides, habitual podcasts, etc. Considering the increasingly uncritical deference to LLM outputs since 2022, Dede (2025, May 19) pointed out the risk that is often unconscious to many of us in a single, clear sentence that is cited in the beginning of this brief review article.

Building on the highlight of LLMs limitations, Dede (2022, December 6) not only draws our attention to the significant value of human IA to remedy the deficiency of AI, but also pin-points the essence of the LLMs roles in classroom practices and educational research with a rigorous analogy of the microscope and telescope: Only when humans use microscopes to figure out where the rich data is, how to collect and analyze the data, can the microscopes change everything and brings new insights into the data. It is the same case with human IA via AI as a tool. Therefore, it is how well we humans use GenAI in

classroom practices that will bring forth our transformational insights in education. So the implication is, without the human efficient and innovative use of the tools such as GenAI agents in classroom practices for human IA, GenAI alone won't produce much creative transformative effect for this purpose, and often turn into digital stopgaps to trap us in an outdated industrial-time educational framework from the past.

Lastly, Dede (2022, December 6) made a bold prediction on the possible elevated human achievements with IA in the next decade in the following three areas:

1. evidence-based modeling for some human decisions
2. higher threshold for human creation
3. better understanding of self and others

Bridging the gap in Learning engineering principles stemmed from IA

Educational researchers are confronted with a major challenge: how to produce a successful research plan in learning engineering for IA to secure funding? Dede (2025, May 24) defined one common pitfall for this challenge: less innovational research proposals often exhibit narrow applicability, manifest effectiveness only within a restricted range of specific contexts, and involve a limited population, and contingent upon highly particularized conditions. This does not help the front-line teachers and schools to understand if and how to implement it in specific classroom practices. This may explain why billions of dollars were spent in annual education funding on reading classrooms in the past decades but not much substantial improvement in K-12 literacy classroom practices and outcomes.

To help fix this ambiguity and disconnection problem, Dede (2025, May 24) proposed a minimalist 3-W method to assembly three essential structural parts of evidence in learning engineering, aiming to clarify the interactional dynamics of the essential gears within the schematic diagram of a desirable research proposal in learning engineering—what works for whom in what conditions, as visualized here as a 3-W schematic diagram in Figure 1. What here indicates what purposes and content of the innovation tasks or actions at hand. For whom defines the scope, target, levels of the tasks, and the implication for the specific standards involved. What conditions delineates the context in which the innovation will work. Only when these three essential gears of learning engineering are molded properly, fit together efficiently, and move in the right direction(s), can this mechanism work well and smooth to fulfill its purpose.

Figure 1.

The 3-W schematic diagram in learning engineering research (Dede, 2025, May 24)



Besides that, building on the diagram of learning engineering process developed by Kessler, Craig, Goodell, Kurzweil and Greenwald (2022), Dede (2025, May 24) outlined four sets of questions to concretize the practice-oriented connection between learning engineering process and classroom practices:

1. What questions do we want the data to answer?

2. What data will answer our questions? What instrumentation do we need to collect it?
3. Are we getting the data we need?
4. What do the data tell us? What other data could support iterative improvement?

Next, to bridging classroom practices and research methodology via learning engineering for human IA, Dede (2025, May 24) described two examples of crucial innovations in research methodology in learning engineering in AI age: Design-based Implementation Research (DBIR) and Component-Based Research (CBR). He also discussed how big data in AI age can empower this aspirational opportunity with the new research methods in this keynote speech.

Moreover, Dede (2025, May 24) introduced some relevant and useful further references such as the guiding principles and typical effects in learning engineering design from Koedinger and Anderson (2013), epistemological considerations about the core tenet of constructivist and socio-cultural theories of learning for UX design in learning engineering (Kolodner, 1993, 2003), theoretical contexts and design-based examples of learning engineering for online education (Dede, et al., 2019) and instructional design and practical learning engineering toolkit (Goodell & Kolodner, 2023).

Bridging the gap in implications for classroom design & assessment with GenAI

Building on the differences between humans and machines, Dede (2022, 2023) noted its implication on human IA that AI will not replace those human skills that AI is not good at, such as jobs requiring some delicate manual work, socio-emotions, or advanced thinking such as high-quality critical thinking and creativity, etc.

According to Dede (2023, August 7), the priority thing in upskilling for human IA is our purpose-setting change in learning in GenAI age. The implication for classroom practice is that, first of all, both educators and learners should not focus on bettering our technology use in classrooms, but on bettering our learning. Secondly, the outcome artifacts such as our lines of coding or written essays, now represent only a proxy or at most a partial destination for our learning in GenAI age. The ultimate destination of our learning journey, i.e., our paramount purpose of learning in GenAI age in the case of scientific writing, is the advanced thinking skills developed during the learning journey itself. The purpose of writing an essay is not just to complete an assignment, but to practice the skill of communicating your own ideas effectively in a written format (Dede, 2023, August 7)

Dede (2023, August 7) took the essay writing process for example: the essay artifact itself is only a proxy for mastery learning, not the ultimate destination of the learning journey. The learners are not likely to achieve the real mastery learning unless they fulfill two pre-conditions in GenAI age: 1) they write, evaluate, rewrite, re-evaluate, rethink and reorganize repeatedly; 2) they consciously refuse to let anyone or any autocomplete technology including GenAI to do their thinking for them, though they may have mechanical assistance in data collection or knowledge searching from GenAI. This reveals an important yet often overlooked pitfall for the AIED developers and AI companies. When the tech companies develop AIED products such as AI-assisted learning analytics or tutoring systems without much awareness of these preconditions for mastery learning, the device may bluntly tell the learners the answers directly or limit the learners' trial efforts to once only, unconsciously depriving the learners' rights and space to develop their own thinking skills independently through repeated rounds of trying and failing in learning.

This shift in epistemological purpose for human IA in learning also has some significant implications for a corresponding change in epistemology for educational assessment in GenAI age. Firstly, we need to assess not only the learning outcome artifacts such as how well the python coding or written essays are, but also how well is the learners' advanced thinking (E.g., how well is their critical or creative thinking exhibited), metacognitive skills or some other uniquely human skills, displayed during the learning process. Secondly, given the fact that the gigantic AI database outwits any single human memory for sure, it is necessary to shift the testing focus from assessing the students' knowledge memory competency to assessing their advanced thinking competency and other uniquely human competency or skills that AI is not good at. Thirdly, GenAI is increasingly doing well in quantitative learning and testing such as the conventional psychometrics testing from last century.

Therefore, if we continue to use the traditional psychometrics testing in GenAI age, we are actually preparing our students to be defeated by GenAI. This will inevitably further devalue the school education by aggravating cheating and dropping-outs issues in AI age. In this case it is high time that we adapt the conventional psychometrics assessment to fit the new demands of assessment in GenAI age.

Secondly, Dede (2022, December 6) proposed multiple dimensions to personalize learning for human IA. Four major dimensions of evidence for personalized learning are synthesized here for reference of classroom practices:

1. Environmental dimensions such as time, place, learning pace, student background, school execution, etc.
2. Learners' self-regulation dimensions including students' choice in learning, attitudes, voice, relationship, etc.
3. Social-emotional learning.
4. Cognitive learning.

Next, Dede (2025, May 19) showed us one example of augmenting creativity with GenAI. He firstly defined creativity as ability in generating innovative and valuable outputs by moving beyond existing knowledge. The significant and common pitfalls include our misbelief in GenAI creativity and undermining the necessity of us humans learning languages. This potentially allows automatic tools like GenAI or Google Translate to take over human writing tasks, displacing human linguistic and literacy development. Apart from highlighting the potential risks, Dede (2025, May 19) emphasized the importance and necessity of languages learning and writing for human IA: Languages are in essence our ways of thinking. Learning is a complex psychological process via language (Goodell, et al., 2023). Languages are in fact the unique and creative products of human culture and context since no other animals on the Earth ever develop languages systems as we humans do. Therefore, we cannot afford to abandon our unique ways of thinking and learning via languages or allow automatic tools like GenAI to replace our in-born uniquely human linguistic competency in learning and using languages, change our cultural heritage in languages, or deprive our human ways of thinking and creation via language.

Fourthly, Dede (2023, 2025) pointed out the nature and limitations of LLM creativity with analogies of the moonlight reflection rather than producing the real light itself: LLMs have no subjective experiences, self-awareness, personal autonomy, physical existence, relationships, cultural background, and moral values (Dede, 2023, August 7)

Building on these issues, Dede (2025, May 19) illustrated some good uses of GenAI in fostering human creativity by some major organizations and institutions as well as the risky uses of GenAI in developing human creativity via several metaphors. Furthermore, he proposed a simple yet insightful solution in augmenting creativity: The essence of creativity lies not just in improving existing processes ("doing things better"), but in using tools like AI to pursue entirely new and more valuable endeavors ("doing better things") (Dede, 2025, May 19).

Conclusion

In summary, this paper synthesized three core pillars of epistemology framework on IA are to identify and bridge the gaps in epistemological foundations of IA, learning engineering principles stemmed from IA, and implications for classroom design & assessment with GenAI. As is discussed, learning is a complicated psychological process. It is necessary for us teachers, learners, leaders, learning engineers and researchers to further consider relevant issues such as What should educators do differently with GenAI in comparison of education in 20th century for human IA? How should learning engineers design GenAI tools that do not pass down the old unfit duct-tape from the 20th century? What new research challenges does human IA raise for the field of learning engineering?

Acknowledgement

The author wishes to express sincere gratitude to Prof. Danielle McNamara for her very thoughtful and constructive feedback throughout the review process. Prof. McNamara's careful attention to clarity, coherence, and methodological rigor substantially

improved the quality and presentation of this article.

References

- Dede, C. (2022, December 6). Intelligence Augmentation via Artificial Intelligence: IA rather than AI. <http://www.youtube.com/watch?v=GFGAQvnlhQk>
- Dede, C. (2023, August 7). Keynote: If AI is the Answer, What is the Question: Thinking about Learning and Vice Versa. <https://www.youtube.com/watch?v=1yZbUwcVR8k>
- Dede, C. (2024, November 11). Complementing Human Minds with Digital Brains: The Role of GenAI in Learning. <http://www.youtube.com/watch?v=WWCP0raajO0>
- Dede, C. (2025, April 12). Understanding GenAI through Mythological Metaphors. http://www.youtube.com/watch?v=dmwqZ_L1yng
- Dede, C. (2025, May 19). Augmenting Human Creativity with Generative AI. https://www.youtube.com/watch?v=HrNY_FFUQpw
- Dede, C. (2025, May 24) (06:08:12: What works when for whom under what conditions: How AI and LE can empower Design-based Implementation Research (DBIR) and Component-Based Research (CBR). <https://www.youtube.com/watch?v=VgHawAWyu6U>
- Dede, C., Richards, J., & Saxberg, B. (Eds.). (2019). Learning engineering for online education. New York, USA: Routledge.
- Georgetown University Center on Education and the Workforce. (2020). Workplace basics: The competencies employers want. Georgetown University.
- Goodell, J., & Kolodner, J. (Eds.). (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Routledge.
- Ginsberg, R., & Zhao, Y. (2025). Reconsidering literacy in an AI world. *Phi Delta Kappan*, 106(7-8), 44-47.
- Kessler, A., Craig, S. D., Goodell, J., Kurzweil, D., & Greenwald, S. W. (2023). Learning engineering is a process. In *Learning Engineering Toolkit*. Taylor & Francis.
- Koedinger, K. R., & Anderson, J. R. (2013). Reifying implicit planning in geometry: Guidelines for model-based intelligent tutoring system design. In *Computers as cognitive tools* (pp. 15-46). Routledge.
- Kolodner, J.L. (1993). *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann.
- Kolodner, J. L., Gray, J., & Fasse, B. B. (2003). Promoting transfer through case-based reasoning: Rituals and practices in learning by design classrooms. *Cognitive Science Quarterly*, 3(2), 183-232.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.
- Thai, K. P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner J. (2023). Learning Engineering is Human-Centered. In J Goodell (Ed.), *The Learning Engineering Toolkit* (pp. 83-124). Routledge.



Charm-bots: The Impact of A.I.'s Sycophancy Language on User Trust

Mercedes Molina, Heather C. Lum, Selena Evans, Laura H. Inderberg, Isabel Glass, & Allie Michael

AI in Education

Artificial Intelligence

chatbots

Large Language Models

student-AI interaction

sycophancy

tone

trust calibration

With the rapid advancement of artificial intelligence (AI) technologies, particularly Large Language Models (LLMs) like ChatGPT, it has become increasingly important for researchers to explore how these innovations influence human-technology interactions, education, and broader societal dynamics. The language used by chatbots can shape how users engage with them. This study aims to examine how the “tone” of AI affects users’ perceptions of chatbots and their interactions with them. By comparing interactions in two contexts: objective (e.g., solving chemistry problems with chatbot assistance) and subjective (e.g., seeking advice for medical symptoms), the research also investigates whether context moderates these interactions. Ultimately, this study contributes to the growing body of knowledge on the factors influencing human-AI interaction and seeks to inform the design of educational AI systems that promote appropriately calibrated trust and support productive learning.

Introduction

As society heads towards an increasingly automated future, its implication on humanity becomes a significant area of focus. One sector of particular interest is human-robot interaction (HRI) research, or the study of human-robot communication—the exchange of information (Goodrich & Schultz, 2007) and, increasingly, student–AI interaction within educational technology and learning systems (e.g. intelligent tutoring systems) (Baker et al., 2022). Large language models (LLM), such as ChatGPT and Google’s Gemini, have captured public attention, with students and the common person using these chatbots for various tasks (Chatterji et al., 2025). The literature on such advanced systems—which produce human-like responses based on large datasets of information—suggests that people do interact differently with chatbots than they do with other humans.

Amongst learning and decision making, subsequent behavior towards automation such as reliance is a significant topic of discussion in the context of understanding how HRI guides human behavior. Research indicates that trust is a crucial factor in predicting reliance on robotic systems, demonstrating a positive relationship (Hoff & Bashir, 2015; Lee & See, 2004). While trust and reliance research in AI is saturated, there remains a gap in work examining the interaction between students and learning systems. The design question must be reframed—not whether trust is higher, but whether student trust is calibrated and appropriate to the task context (e.g. when students rely on the AI and when they choose to seek a second opinion). However, there are caveats to this relationship, such that trust and reliance in AI can be rapidly restored after early and late incidence of error, although early errors have a much more significant negative impact on reliance compared to later errors made (Kahr et al., 2024). Shahrasbi (2025) proposes a theoretical experimental design in which participants interact with a

neutral tone agent and an empathetic tone agent, resulting in a preference for the empathetic tone agent for future use, which they identify as a behavioral measure of trust. They build on this by suggesting a second experiment, where individuals would complete a multi-step task with the option of switching between the two agents varying in tone after each response, hypothesizing that participants will be more tolerant of an erroneous empathetic agent if they began the task with it, thus providing an established bond.

The current study at hand serves to address the research gap on how chatbot tone and subjective versus objective contextual questions influence subsequent student trust calibration and help-seeking/continued-use intention toward an LLM-based AI, while investigating user attribute interaction effects. As chatbots increasingly appear in learning environments such as tutoring and educational decision-support, understanding how tone interacts with question type is directly relevant to inform learning system design. To our knowledge, the proposed framework is novel within the literature.

Methods

Measures

Participants will complete a post-study survey that includes demographic information and measures of GenAI usage, prior exposure and experience, and fairness-related questions. To capture individual differences relevant to technology adoption and responses to AI, established scales are included, such as the Technology Readiness Index (TRI) (Parasuraman, 2000), the Big Five Personality Dimensions (Gosling et al., 2003), the Maranell Religiosity Scale (Maranell, 1974), and the AI Anxiety Scale (Wang & Wang, 2019).

Procedure

A between-subjects design will be used, where a total of 34 participants (students) randomly assigned to one of two groups. One group will interact with the LLM, ChatGPT, which will be personalized to use a “friendly” tone ($n = 17$), while the other group will interact with the same chatbot prompted to maintain a “neutral” tone ($n = 17$).

Once participants arrive at the study site, they will be given a consent form to read and fill out in order to proceed with the study. After participants provide their consent, they will interact with ChatGPT as a proxy for a chat-based AI tutor interface that could appear in a learning platform.

Participants will be tasked to ask ChatGPT objective questions; help understanding chemistry concepts (e.g. “What’s benzene’s most distinctive feature?”) and subjective questions; advice on vague medical symptoms (e.g. You lost your appetite since the fall). This will allow us to analyze how interactions differ between interpreting confusing course topics and engaging with other decision-support prompts where the “best” answer is subjective rather than strictly correct. After finishing the chatbot interaction, participants will complete a short series of questionnaires in Qualtrics on generative AI usage and preferences, personality test questions, technology acceptance, and demographics questions.

Results

This study investigates the influence of a chatbot’s language style on user trust and the potential moderating role of individual differences. We hypothesize that the use of sycophancy language by the chatbot will increase students’ self-reported trust and intention to continue using the chatbot, with trust being higher in objective domains than in subjective ones. This objective–subjective contrast is intended to surface whether trust is calibrated to question type. Additionally, we will examine whether individual characteristics affect trust, including familiarity with generative AI (genAI) technologies, attitudes toward new

technologies, AI-related anxiety, and personality traits such as extraversion, agreeableness, conscientiousness, and openness to experience. Specifically, we expect that users with greater familiarity with genAI technologies and more positive attitudes toward new technologies will exhibit higher trust, whereas those with higher AI-related anxiety will exhibit lower trust.

Discussion

If the hypotheses of this study are supported, the findings could have far-reaching implications for the design, deployment, and ethical governance of large language models (LLMs) in educational technology and learning systems (e.g., AI tutors). Demonstrating that a sycophancy tone increases user trust would suggest that developers could strategically use conversational style to enhance engagement and repeated use of AI systems in student–AI interaction. At the same time, such effects raise important concerns regarding transparency, informed user consent, and the potential for overreliance on AI-generated information. Dependence on AI could be particularly problematic in educational decision-making contexts, where reliance on inaccurate or misleading outputs could result in tangible consequences. Moreover, if individual differences such as familiarity with technology, personality traits, or AI-related anxiety are shown to influence trust, these findings could inform the development of adaptive AI systems that calibrate their communication style according to user characteristics, thereby balancing usability with caution and supporting productive learning. From an ethical standpoint, these results highlight the necessity of incorporating safeguards to prevent manipulation or unintended biases from user behavior. They also underscore the importance of promoting AI literacy and public awareness to ensure that users approach AI outputs with an appropriate level of critical evaluation. Ultimately, such insights contribute to a more responsible, user-centered approach to AI deployment, emphasizing both the benefits and ethical responsibilities associated with advanced generative technologies in learning engineering and learning system design.

References

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*. Advance online publication. <https://doi.org/10.1037/tmb0000058>
- Goodrich, M. A., & Schultz, A. C. (2008). Human–robot interaction: A survey. *Foundations and Trends® in Human–Computer Interaction*, 1(3), 203–275.
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the big five personality domains. *Journal of Research in Personality*, 37, 504–528.
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3), 407–434.
- Kahr, P. K., Rooks, G., Snijders, C., & Willemsen, M. C. (2024, March). The Trust Recovery Journey: The Effect of Timing of Errors on the Willingness to Follow AI Advice. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (pp. 609–622).
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50–80.
- Maranell, G. M. (1974). *Responses to religion: Studies in the social psychology of religious belief*. Lawrence, KS: University Press of Kansas.
- Parasuraman, A. (2000). Technology Readiness Index (TRI): A multiple-item scale to measure readiness to embrace new technologies. *Journal of Service Research*, 2(4), 307–320.

Shahrashbi, Nasser, "Empathy in AI Agents: Impacts on User Trust and Error Tolerance" (2025). AMCIS 2025 TREOs. 119.

Wang, Y. Y., & Wang, Y. S. (2019). Development and validation of an artificial intelligence anxiety scale: an initial application in predicting motivated learning behavior. *Interactive Learning Environments*, 30(4), 619–634.
<https://doi.org/10.1080/10494820.2019.1674887>



Coaching, Not Autocomplete: Early Evidence from ConnectInk's AI-Supported Personal Narrative Pilot

Julio Intriago-Izquierdo & Rahul Patel

AI Writing Coach

Feedback

learning engineering

Personal Narrative

Student Voice

This short paper reports early findings from a Spring 2025 classroom pilot of ConnectInk, a browser-based, guardrailed AI writing coach that supports, but never generates, student text. We frame the pilot as a nested Learning Engineering (LE) cycle within ConnectInk’s larger product-and-partnership effort. In partnership with the Center for the Professional Education of Teachers at Teachers College, Columbia University, the tool was embedded in a brief personal narrative unit across three NYC high schools. A mixed-methods design (baseline perception survey, post-lesson reflections, observation logs, and pre/post writing samples) examined perceptions, process, and products. Students entered with low enjoyment and limited sharing comfort but valued feedback and showed emerging openness to AI for revision. Across the unit, students used ConnectInk to brainstorm, organize, and revise; paired samples and commentary noted growth in narrative craft, structure, and elaboration. Findings suggest that bounded AI coaching, coupled with genre-based pedagogy, can feasibly support confidence, craft, and audience awareness within a brief instructional arc.

Julio Intriago-Izquierdo¹ and Rahul Patel²

¹ Teachers College, Columbia University; jei2116@tc.columbia.edu

² Insights & Innovations LLC; rahul@innovateinsightfully.com

Abstract.

Keywords: Learning Engineering, AI Writing Coach, Personal Narrative, Student Voice, Feedback.

Introduction

Many students, particularly in large urban systems, struggle to see themselves as writers and to receive the timely, individualized feedback that builds skill and motivation. In New York State, 54% of students fall short of grade-level expectations in ELA (NYSED, 2024). Evidence syntheses underscore feedback’s large effects on learning (Hattie, 2008). In addition, research on adolescent writing instruction emphasizes the importance of explicit strategy support and iterative practice, and scholarship on writing and identity highlights how writing can function as an act of self-positioning and voice (Graham & Perin, 2007; Ivanič, 1998).

Against this backdrop, we developed ConnectInk, a tool designed to expand access to feedback while protecting student voice. ConnectInk, developed by Insights & Innovations with the Center for the Professional Education of Teachers (CPET) at Teachers College, Columbia University, embeds a guardrailed, question-driven AI coach inside a genre-based narrative writing unit. Crucially, the AI never drafts text; it only offers prompts, genre-specific cues, and audience-awareness questions, preserving student authorship.

Guided by this theory of change, this short paper addresses three research questions (RQs): RQ1: How do students perceive the drafting and revision experience when supported by an AI-powered tool that does not generate content for them? RQ2: How does writing a personal narrative, supported by an AI-powered tool, shape students' self-awareness and sense of identity? RQ3: To what extent do shared narratives influence peer understanding and audience empathy?

Learning Engineering lens. Learning Engineering (LE) emphasizes the systematic, human-centered, and evidence-driven design of learning experiences through iterative cycles that connect learning science, design decisions, and data. In this study, we treat the Spring 2025 pilot as a nested LE cycle inside ConnectInk's larger product-and-partnership effort: the instructional unit and tool behaviors operationalize theory-informed design decisions, the classroom implementation tests feasibility under real constraints, and the mixed-methods data plan produces actionable evidence to refine both the learning experience and the tool (Baker et al., 2022; Goodell et al., 2023; Thai et al., 2022; Craig et al., 2025).

Methods

Design

CPET led a pre/post mixed-methods study approved by the Teachers College IRB (Protocol "ConnectInk: Storytelling & Community," #25-325; issued 04/15/2025). The intervention consisted of a six-lesson personal narrative unit typically implemented over 4–6 class sessions (not all classes completed all six lessons), integrated with ConnectInk as an in-class writing coach, and facilitated by CPET writing coaches. Across sessions, students followed a consistent instructional arc—brainstorm → draft → revise → share—with ConnectInk used at defined moments to prompt idea generation, support drafting momentum, and guide revision without generating student text.

Participants & Setting

Three Title I New York City high schools participated: Northside Charter High School (Brooklyn), Leaders High School (Brooklyn), and East Bronx Academy for the Future (Bronx). Two coaches implemented the unit with 354 students.

Intervention

The intervention combined a genre-based narrative writing sequence with ConnectInk (<https://connectink.org>), a browser-based, guardrailed AI writing coach. Students used the tool through multiple entry points: (a) low-stakes creative prompts and mentor texts; (b) "Get Unstuck" nudges during drafting; (c) genre-based feedback during revision; (d) audience-aware prompts to support clarity and purpose; and (e) a review-for-publication checklist to prepare final drafts. Guardrails ensured the AI asked questions and offered cues but never produced student text.

Measures

Data sources included: (1) an 18-item Student Perception Survey administered at baseline; (2) post-lesson student reflections; (3) pre/post narrative writing samples; and (4) coach observation logs. In total, 217 narratives were drafted; 188 pre-writing samples and 29 post-project samples were collected for text analysis.

Analysis. Survey items were summarized descriptively. Qualitative data (student reflections, coach logs, and writing samples) were thematically analyzed, and CPET literacy experts reviewed exemplar paired samples to identify changes in narrative craft and structure. From a Learning Engineering perspective, this cycle used an explicit measurement plan spanning perception (survey), process (reflections/logs), and product (writing samples) to connect design intent to observed learner experience and outcomes and to specify iteration targets for the next cycle (Totino & Kessler, 2024).

Results

RQ1: Writing Process with AI Coaching

At baseline, 37% of students reported enjoying writing and 25% felt comfortable sharing work, while 68% valued receiving feedback. AI use was emerging, as 44% had used AI for writing, 36% for feedback/revision, and 36% believed AI could help them improve (24% disagreed). Reflections and coach logs suggest students most often used ConnectLink to get started and revise early craft moves, supporting brainstorming, organization, and revision toward stronger hooks, clearer structure, and added detail. For example, one student wrote, “I used the AI tool to pick my prompt on what to write and also to see if my hook was good.” Coaches similarly noted that students “liked that the tool prompted them and helped them see what their writing did,” and multiple students emphasized they were glad “it didn’t generate things for them.”

RQ2: Self-Awareness & Writer Identity

Baseline identity indicators were modest, as 37% felt confident in their writing and 35% saw themselves as good writers, though 63% agreed they were learning to write better. Reflections showed increased self-diagnosis (e.g., noticing needs in organization and elaboration) and surprise at one’s capacity to write more than expected. One student captured this shift directly: “I learned that i am not a bad writer when writing about something im interested in.” Coach logs echoed this pattern; in one case, a student was “particularly surprised by the feedback” because it surfaced an interpretation of their narrative they “hadn’t considered,” suggesting the tool may support metacognitive insight about voice and meaning beyond surface-level edits.

RQ3: Peer Connection & Audience Empathy

Students reported that writing helped them understand themselves (43%) and that reading peers’ work improved understanding (46%) and sometimes changed opinions (43%). Reflections indicate early signs of perspective-taking, including noticing classmates’ values through chosen objects and story details, as well as learning that “People have different points of view” and “How others see things differently than me.” Coaches observed partner sharing marked by “smiling and nodding... and eye contact that indicated focused listening,” while also noting hesitancy to share narratives with the full group, an important constraint on empathy-building via wider audience exposure.

Writing Growth Evidence

Paired samples showed clear gains from brief, expository-leaning responses to extended narratives with recognizable arcs, vivid hooks, dialogue, internal monologue, and reflective endings. CPET expert commentary highlighted shifts toward “show, don’t tell,” rhetorical questioning, and elaboration. All final texts were 100% student-authored due to non-generative guardrails.

Discussion

Within a short, feasible arc, combining a guardrailed AI coach with genre-based pedagogy yielded promising movement in confidence, craft, and classroom connection. The findings align with evidence that writing improves when learners receive explicit strategy support, iterative practice, and actionable feedback (Graham & Perin, 2007; Hattie, 2008), and with scholarship emphasizing writing as a site of voice and identity work (Ivanič, 1998). Practically, the approach shifts AI from text generator to dialogue partner, preserving authorship and ethics while helping educators concentrate time on higher-order instruction, conferencing, and community-building in high-need settings (NYSED, 2024).

Limitations include end-of-year timing (many classes completed 4–5 of 6 lessons), partial post-intervention survey completion, and early-stage product maturity. Next steps include implementing earlier in the semester to complete all six lessons, tightening post-lesson reflections to capture clearer evidence of writing decisions (e.g., what prompt was used and what change was made), addressing high-friction usability issues that interrupt drafting/revision, and completing the planned mixed-methods analysis.

In this pilot, student reflections and coach logs provided high-signal evidence about where coaching prompts supported momentum and revision, and where access/workflow barriers constrained uptake, making iteration targets concrete. This work demonstrates a transferable nested Learning Engineering cycle for AI-supported writing instruction, in which teams: (1) translate learning goals and constraints (student voice, authorship, narrative craft) into explicit design requirements (non-generative guardrails; question-driven coaching); (2) embed the tool in a defined instructional sequence with roles for teachers and peers; (3) implement under authentic constraints; (4) instrument for evidence across perception, process, and product; and (5) use findings to specify iteration targets for the next cycle. Framing AI as a bounded dialogue partner, rather than an autocompletion engine, supports ethical authorship while producing actionable design knowledge for scaling voice-centered writing experiences.

Acknowledgments

We thank Dr. Roberta Lenger Kang, Dr. Cristina Compton, Dr. Kelsey Hammond, and Gregory Petershach of the Center for the Professional Education of Teachers (CPET) at Teachers College, Columbia University, for partnership and expert review; Joseph Martinez (Insights & Innovations) for software development; and the participating teachers and students at Northside Charter High School, Leaders High School, and East Bronx Academy for the Future for their contributions to implementation and data collection. This work was supported by The Wellness Classroom.

References

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*, 3(1), 1–23. <https://doi.org/10.1037/tmb0000058>
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale*. <https://doi.org/10.59668/2109.21735>
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance: Based on the iFEST poster (Winner of Best Poster Design). *Journal of Military Learning*, 7(1).
- Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools – A report to Carnegie Corporation of New York*. Washington, DC:Alliance for Excellent Education.

Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement* (1st ed.). Routledge.
<https://doi.org/10.4324/9780203887332>

Ivanič, R. (1998). *Writing and identity: The discorsal construction of identity in academic writing*. John Benjamins.
<https://doi.org/10.1075/swll.5>

New York State Education Department. (2024). State Education Department releases preliminary data on ELA, Mathematics, and Science exams. <https://www.nysed.gov/news/2024/state-education-department-releases-preliminary-data-english-language-arts-mathematics-and-science-exams>

Thai, K.-P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. (2022). Learning engineering is human-centered. In J. Goodell & J. Kolodner (Eds.), *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond* (pp. 83–123). Routledge. <https://doi.org/10.4324/9781003276579-7>

Totino, L., & Kessler, A. (2024). “Why did we do that?” A systematic approach to tracking decisions in the design and iteration of learning experiences. *The Journal of Applied Instructional Design*, 13(2), 277–283. <https://doi.org/10.59668/1269.15630>



Comparing Epistemic Emotions and User Experience Across Two AI Instructional Designs in Biology Learning

Yiwen Li, Ying-Chih Chen, Christopher Day, & Ella Tour

AI-supported learning

Epistemic Emotions

Science Education

This pilot study investigates how two distinct AI instructional designs shape undergraduate students' epistemic emotions and user experience in biology education. Students interacted with either an AI-Tutor that provided structured conceptual guidance or an AI-Navigator that incorporated an uncertainty-centered instructional approach designed to strategically raise, maintain, and reduce conceptual, procedural, and epistemic uncertainty. After completing a biology lab quiz, students completed the Epistemically-Related Emotions Scales (EES) and several user-experience items. AI-Tutor users reported significantly higher enjoyment, perceived helpfulness, AI preference, and future use intention; while AI-Navigator users showed more varied epistemic-emotion profiles, including slightly elevated confusion. Drawing on Control–Value Theory and the ERAS (Emotion Regulation in Achievement Situations) model, these findings highlight a design tension, which is that systems that enhance perceived control tend to produce more positive epistemic emotions and user experiences, compared to the systems that introduce epistemic friction in the short term.

Introduction

AI-based instructional tools are becoming increasingly common in biology and science education, where they support students as they interpret evidence, solve problems, and construct scientific explanations (Aripin et al., 2024). Recent work shows that generative AI systems do more than streamline tasks. They also shape learners' cognition, epistemic engagement, and emotional experiences as they make sense of complex scientific ideas (Yan et al., 2025). Epistemic emotions play a central role in this process because they signal impasses, motivate deeper inquiry, guide metacognitive regulation, and influence students' willingness to persist with challenging problems (Muis, Chevrier, & Singh, 2018). However, despite the growing use of AI in science classrooms, much remains unknown about how specific AI design features shape the emotional dynamics that arise during scientific reasoning, particularly emotions linked to uncertainty such as curiosity, confusion, surprise, and frustration.

Biology learning frequently involves scientific uncertainty, whether students are analyzing data, evaluating competing interpretations, or reasoning through complex processes (Chen, Park, & Rapkiewicz, 2024). These moments often evoke epistemic emotions, such as confusion, surprise, or curiosity, that influence how learners process conflicting information, regulate their thinking, and engage with the task (Pekrun et al., 2017). The Control–Value Theory of emotions explains that these emotions arise from learners' appraisals of control ("Can I handle this uncertainty?") and value ("Does resolving this

matter?”) (Pekrun, 2024). High perceived control typically elicits enjoyment and curiosity, whereas reduced control, which is common in uncertainty-rich situations, tends to evoke confusion or frustration (Pekrun, 2024).

The ERAS extends this perspective by describing how learners regulate emotions during academic tasks. ERAS integrates the principles of Control–Value Theory with Gross’s process model of emotion regulation, explaining how emotions such as confusion emerge from novelty, conflict, or epistemic impasses and how learners reappraise or regulate these emotions (Harley et al., 2019). Research on confusion regulation shows that confusion can support deeper learning when learners successfully regulate it, but it can also impede progress if it overwhelms perceived control (Muis et al., 2025).

This study compares two AI instructional designs that differ in how they structure uncertainty. The AI-Tutor provides structured, clarity-oriented conceptual guidance intended to sustain students’ perceived control. The AI-Navigator builds on the AI-Tutor but embeds the uncertainty-centered instructional approach (Chen et al., 2024), which identifies conceptual, procedural, and epistemic uncertainty and strategically raises, maintains, or reduces that uncertainty to support productive struggle. These designs therefore create different epistemic-emotional experiences. The AI-Tutor may promote smoother, more comfortable interactions, while the AI-Navigator may evoke more effortful but potentially generative emotional states. Using the full 21-item EES (Pekrun et al., 2017) and targeted user-experience items, this study examines how these instructional differences shape students’ epistemic emotions and user experiences during biology problem-solving.

Methods

Participants and Context

Participants were 34 undergraduate students enrolled in a biology laboratory course at a large public university in the southwestern United States. The course had two independently registered lab sections, which created naturally occurring groups for the AI-Tutor and AI-Navigator conditions. Both sections completed the same biology quiz and were instructed to use only their assigned AI tool, without assistance from peers, the internet, or other external resources. Immediately afterward, all students completed a post-quiz survey on epistemic emotions and user experience. Seventeen responses were collected from each section.

Designs of AI-Tutor and AI-Navigator

The AI-Tutor was built using a structured prompt that incorporated students’ background knowledge, key biological concepts, and common misconceptions. It evaluated student responses without providing answers. When students expressed uncertainty or misunderstanding, the AI-Tutor offered a brief conceptual explanation and asked one or two follow-up questions to confirm understanding. This resulted in a structured, clarity-oriented form of guidance.

The AI-Navigator preserved the AI-Tutor’s foundation but added an uncertainty-centered instructional approach. The system identified conceptual uncertainty (gaps in understanding), procedural uncertainty (uncertainty about methods or steps) and epistemic uncertainty (questions about evidence or justification) (Chen et al., 2024). When such uncertainty emerged, the AI-Navigator intentionally increased uncertainty by prompting students to reconsider assumptions, sustained it by withholding confirmation and encouraging deeper reflection, or reduced it when necessary. As with the AI-Tutor, it never provided answers. The design approached uncertainty as a productive space for scientific inquiry.

Measures and Procedure

Students in both groups completed a brief post-quiz survey that included (a) four user-experience items assessing ease of use, perceived helpfulness, preference for AI versus a peer, and intention to use the tool in future biology quizzes; and (b) the

full EES (7 subscales: surprise, curiosity, enjoyment, confusion, anxiety, frustration, and boredom) (Pekrun et al., 2017). The internal consistency reliability estimates were acceptable for all subconstructs ($\alpha = .65-.91$). All items used Likert-type response formats.

Quantitative analyses were conducted in R, including descriptive statistics and inter-correlations among all user-experience and epistemic-emotion variables (Table 3). Group differences between the AI-Tutor and AI-Navigator groups were examined using Welch's t-tests with Cohen's d effect sizes (Tables 1–2). These analyses allowed us to characterize initial emotional and experiential differences across the two AI-supported learning conditions.

Results

Several significant differences in user experiences were found across the two AI designs. As shown in Table 1, students using the AI-Tutor showed a statistically significantly stronger preference for AI over a peer than those using the AI-Navigator, $t(27.89) = -2.32$, $p = .03$, $d = 0.88$. They also demonstrated significantly higher intention to use the tool again, $t(25.13) = -2.41$, $p = .02$, $d = 0.85$, and rated the tool as more helpful, $t(30.92) = -2.35$, $p = .03$, $d = 0.83$. Differences in ease of use favored the AI-Tutor but were not statistically significant ($p = .17$).

Table 1.

Differences in user experience responses between AI-Tutor and AI-Navigator groups.

	Tutor (n=17)		Navigator (n=17)		t	df	p	Cohen's d
	M	SD	M	SD				
AI Preference (vs. Peer)	0.67	0.488	0.27	0.46	-2.32	27.89	0.03	0.88
Future Use Intention	1.82	0.393	1.35	0.70	-2.41	25.13	0.02	0.85
Easiness	3.94	1.03	3.41	1.18	-1.40	31.45	0.17	0.49
Helpfulness	4.12	1.11	3.29	0.92	-2.35	30.92	0.03	0.83

Note. AI Preference (vs. Peer) was coded as Peer = 0 and AI Tutor/AI Navigator = 1. Future Use Intention (FUI) was coded as No = 0, Maybe = 1, and Yes = 2. Easiness and Helpfulness were measured on 5-point Likert scales ranging from 1 (strongly disagree) to 5 (strongly agree).

For epistemic emotions, Table 2 shows that only enjoyment was statistically significantly higher among AI-Tutor users, $t(31.98) = -2.64$, $p = .01$, $d = 0.93$. Other differences showed moderate effect sizes: AI-Tutor users reported more surprise ($d = 0.70$), while AI-Navigator users reported slightly more confusion and boredom without statistical significance.

Table 2.

Differences in epistemic emotions between AI-Tutor and AI-Navigator groups.

Epistemic Emotions	Tutor	Navigator	t	df	p	Cohen's d

	(n=17)		(n=17)					
	M	SD	M	SD				
Surprise	3.25	0.94	2.61	0.96	-1.99	31.99	0.06	0.70
Curiosity	3.69	0.95	3.57	0.68	-0.42	28.81	0.68	0.15
Enjoyment	3.20	1.14	2.18	1.11	-2.64	31.98	0.01	0.93
Confusion	2.69	1.02	2.94	1.00	0.74	31.97	0.47	-0.26
Anxiety	2.59	1.16	2.88	1.17	0.74	31.99	0.47	-0.26
Frustration	2.27	1.10	2.53	1.18	0.65	31.83	0.52	-0.23
Boredom	2.02	0.90	1.98	0.64	-0.15	28.73	0.89	-0.05

Correlation analyses across all participants (Table 3) revealed several statistically significant relationships between epistemic emotions and user experience variables. Positive epistemic emotions, particularly surprise, curiosity, and enjoyment, were strongly and statistically significantly correlated with one another. In contrast, significant positive correlations emerged among negative epistemic emotions such as frustration, anxiety, and boredom. Importantly, future use intention (FUI) and perceived helpfulness showed statistically significant positive associations with enjoyment and surprise, and statistically significant negative associations with frustration. Perceived easiness demonstrated a similar pattern of significant links to both positive and negative emotions. These significant relationships indicate that students' evaluations of the AI tools were closely tied to their emotional experiences: stronger positive emotions predicted higher AI preference and FUI, while stronger negative emotions predicted diminished user experience.

Table 3.

Pearson's correlation matrix for user-experience and epistemic-emotion variables.

Variables	1	2	3	4	5	6	7	8	9	10	11
1. Surprise											
2. Curiosity	0.58*										
3. Enjoyment	0.79*	0.66*									
4. Confusion	0.02	0.16	-0.09								
5. Anxiety	-0.2	-0.03	-0.26	0.69*							
6. Frustration	-0.49*	-0.25	-0.51*	0.69*	0.73*						
7. Boredom	-0.21	-0.29	-0.35*	0.56*	0.56*	0.57*					
8. AI Preference	0.47*	0.1	0.36	-0.26	-0.2	-0.43*	-0.06				

9. FUI	0.52*	0.25	0.49*	-0.23	-0.25	-0.56*	-0.06	0.66*		
10. Easiness	0.35*	0.14	0.28	-0.44*	-0.48*	-0.56*	-0.4*	0.57*	0.46*	
11. Helpfulness	0.58*	0.35*	0.53*	-0.23	-0.33	-0.5*	-0.29	0.58*	0.54*	0.71*

Note. An asterisk (*) indicates a statistically significant correlation ($p < .05$). Higher significance levels ($p < .01$, $p < .001$) are included within this category but are not separately marked.

Discussions

The findings of this pilot study reveal that different AI instructional designs can shape students' experiences in meaningful ways during biology problem-solving. The AI-Tutor produced more positive user experience responses, including higher enjoyment, helpfulness, preference for AI, and future use intention. These findings align with Control–Value Theory, which asserts that enjoyment arises when learners perceive high control. The AI-Tutor's structured and supportive response style likely bolstered perceived control, resulting in more positive emotional reactions.

The AI-Navigator's uncertainty-centered design produced more varied epistemic-emotion profiles, including slightly higher confusion and lower enjoyment. The ERAS model explains that such emotions arise when learners confront novelty, conflict, or ambiguity and emphasizes that the regulation of these emotions is critical for productive learning. Although the AI-Navigator temporarily reduced perceived control, the uncertainty it introduced may support deeper reasoning by prompting learners to evaluate evidence, consider alternatives, or articulate gaps in their understanding.

Correlational findings further highlight the strong relationship between emotional responses and user experience evaluations. Enjoyment and surprise were strongly associated with preference for AI, perceived helpfulness, and future use intention, whereas frustration was negatively associated with these same outcomes. These relationships suggest that students' willingness to adopt AI tools is shaped not only by performance-related factors but also by the emotional climate of their interaction. This dynamic may help explain why the AI-Navigator, although designed to support productive struggle, was evaluated less favorably in the short term.

Together, these findings illustrate a design tension. Systems that reduce epistemic friction create smoother emotional experiences, while those that introduce uncertainty evoke more varied emotional responses. Emotional comfort and epistemic rigor therefore require careful balancing. Future analyses of students' chat logs and conceptual understanding will help determine whether the Navigator's uncertainty-raising moves supported deeper learning, despite lower immediate satisfaction.

References

- Aripin, I., Gaffar, A. A., Jabar, M. B. A., & Yulianti, D1. (2024). Artificial intelligence in biology and learning biology: A literature review. *Jurnal Mangifera Edu*, 8(2), 41-48.
- Chen, Y. C., Benus, M. J., & Hernandez, J. (2019). Managing uncertainty in scientific argumentation. *Science Education*, 103(5), 1235-1276.
- Chen, Y. C., Park, J., & Rapkiewicz, J. G. (2024). Using student uncertainty as a pedagogical resource (SUPeR) approach to support students' productive struggle in the Biology classroom. *The American Biology Teacher*, 86(8), 485-495.

- Chen, Y. C., Jordan, M., Park, J., & Starrett, E. (2024). Navigating student uncertainty for productive struggle: Establishing the importance for and distinguishing types, sources, and desirability of scientific uncertainties. *Science Education*, 108(4), 1099-1133.
- Harley, J. M., Pekrun, R., Taxer, J. L., & Gross, J. J. (2019). Emotion regulation in achievement situations: An integrated model. *Educational psychologist*, 54(2), 106-126.
- Muis, K. R., Chevrier, M., & Singh, C. A. (2018). The role of epistemic emotions in personal epistemology and self-regulated learning. *Educational psychologist*, 53(3), 165-184.
- Muis, K. R., Kohatsu, M., Pekrun, R., & Li, S. (2025). Confusion and confusion regulation: An empirical investigation of the emotion regulation in achievement situations model. *Contemporary Educational Psychology*, 80, 102350.
- Pekrun, R., Vogl, E., Muis, K. R., & Sinatra, G. M. (2017). Measuring emotions during epistemic activities: The epistemically-related emotion scales. *Cognition and Emotion*, 31(6), 1268-1276.
- Pekrun, R. (2024). Control-value theory: From achievement emotion to a general theory of human emotions. *Educational Psychology Review*, 36(3), 83.
- Yan, L., Pammer-Schindler, V., Mills, C., Nguyen, A., & Gašević, D. (2025). Beyond efficiency: Empirical insights on generative AI's impact on cognition, metacognition and epistemic agency in learning. *British Journal of Educational Technology*, 56(5), 1675-1685.



Currents of Inquiry: Insights From Two Years of Real-World AI-Learner Water Conversations

Anqi Shao, Stephen Carradini, Ram Sankar Harikrishnan, Jacob Kuriakose, Suraj Manyalal, & Srinivasan Ravichandran

Conversational AI

Learning Analytics

STEM education

Applying a Learning Engineering framework to informal STEM education, this study investigates the alignment between user inquiry patterns and conversational AI responses. We present an analysis of two years of interactions between our Waterbot and community users asking about water issues. Treating questions as signals of learners' cognitive focus, we code question types and quantify readability, causal connectives, and lexical markers of certainty for both users' questions and the AI's responses. Overall, learning sessions remained short, with learning depth tied more to chatbot verbosity than to linguistic complexity or certainty. These findings offer an initial design baseline for conversational AI that supports users' learning regarding complex issues. The observed asymmetries in length, complexity, and certainty point to concrete levers for the next design cycle, such as briefer default answers and clearer pathways for follow-up questions. As a study grounded in real learner traces, this work establishes an empirical baseline for the next engineering design cycle of conversational

systems aimed at civic and scientific informal learning regarding water issues.

Introduction

Conversational AI has become a common gateway for public information seeking (IPSOS, 2025). Large language models (LLMs) embedded in chat interfaces now function as informal tutors across a wide range of STEM topics. These conversational systems provide real-time interaction as a responsive partner, which may help with engagement and exploration of complex scientific concepts (Kerry et al., 2008). A growing body of research has focused on speculating learning experiences with AI in experimental settings (Carradini, 2024), yet relatively little is known about how real learners actually phrase their questions and how AI systems respond during unstructured, real-world interactions.

Research on AI tutors shows that conversational agents can assist with explanation, feedback, and step-by-step guidance. Recent studies on LLM-based systems indicate that conversational AI can offer coherent explanations, generate examples, and help learners express scientific ideas (Kasneci et al., 2023; Terzimehić, et al., 2025). However, several challenges persist. LLMs often produce lengthy or overly detailed responses (Su et al., 2025) and avoid taking clear positions on uncertain topics (Yona, Aharoni & Geva, 2024). Researchers have also warned about the problem of over-explanation, where educators may present too much information that overwhelm learners or discourage further inquiry (Fleisher et al., 2025). While many controlled experiments test AI tutors on structured tasks, far less is known about how people outside formal education settings use conversational AI when exploring scientific topics. Informal interactions may involve brief exchanges, loosely formed questions, and variable levels of engagement. Therefore, a challenge remains concerning misalignment between the efficiency-oriented nature of informal information seeking and the sustained cognitive effort required for deep scientific learning. This creates a potential pedagogical mismatch in which system outputs exceed the learner's capacity for productive engagement. Addressing this issue requires close analysis of how such interactions unfold in actual use contexts.

This study addresses these gaps by analyzing about 3,000 real exchanges between users and Waterbot, a water-related conversational system (Rajan, Carradini & Lauer, 2024). The Arizona Waterbot is a community-engaged artificial intelligence project that applies learning engineering principles to environmental communication. Developed in 2023, Waterbot provides clear, context-specific answers about Arizona's water management, conservation, and drought conditions. In addition to its technical structure, Waterbot serves as a pedagogical tool that supports civic learning by helping residents understand scientific and policy information, as well as identify practical ways to promote local sustainability. By analyzing user-chatbot interactions up to July 2025, the project reflects on connecting data, design, and educational science to create equitable, transparent, and human-centered learning solutions. Drawing on principles of learning engineering, the analysis examines linguistic and conversational features and how these elements correspond to AI-mediated learning about environmental issues. By studying authentic interactions, this pilot research offers an empirical basis for improving conversational design and promoting more effective inquiry in future learning systems.

This study frames Waterbot as a learning engineering system, developed to improve learner outcomes through data-driven iteration (Baker et al., 2022; Goodell, Kessler, & Schatz, 2023). While the full learning engineering cycle involves continuous refinement from problem definition, design, data collection, analysis, and redesign (Goodell, Kolodner & Kessler, 2022), this paper specifically reports on one analytic cycle embedded within the longer design trajectory. By examining authentic interactions from the system's initial deployment, we aim to generate evidence that informs the subsequent redesign phase. This approach ensures the AI evolves through human-centered engineering design to better support public engagement with complex environmental topics (Thai et al., 2022).

Methods

To systematically isolate and improve interaction dynamics, this analysis functions as a nested learning engineering cycle within the broader Waterbot design trajectory (Craig et al., 2025). This specific cycle targets the learning problem wherein informal civic learners often engage with complex environmental topics through brief, potentially fragmented interactions. We investigate the design assumption that specific system features may affect the learner's ability to sustain inquiry. To test this, we examine analytic evidence derived from linguistic traces of real-world use, allowing us to track how design decisions influence learner engagement (Totino & Kessler, 2024).

Data source

The dataset consisted of 20,779 raw user-AI interaction logs collected from Waterbot. Initial preprocessing involved removing duplicated entries and unusable inputs such as garbled machine-generated text, resulting in 4,595 cleaned records. To ensure analytic consistency, only messages automatically identified as English-language were retained. This yielded a final analytic dataset of 2,978 user queries paired with the chatbot's responses. Each interaction was tagged with a system-generated conversation ID, allowing individual messages to be grouped into conversational sessions.

The primary units of analysis were individual message pairs, consisting of (a) a single user query and (b) the chatbot's immediate response. All computations of linguistic features (e.g., readability, certainty, causal expressions) were performed separately for users and the chatbot. Conversation-level analyses used the conversation ID to aggregate message-level features. All analyses were performed in Python 3.11 using standard NLP and statistical libraries.

Key Variables and Operationalization

Question Type.

User inputs were categorized according to their interrogative form including what, how, why, who, where, when, yes/no, and non-question prompts. Classification was rule-based, using the first interrogative token or syntactic structure of the message. This variable informed analyses of inquiry depth and the distribution of surface- versus mechanism-oriented questions. Theoretically, question phrasing serves as the learner's entry signal (Chin & Osborne, 2006), distinguishing between users seeking static facts (surface learning) and those attempting to construct mental models (deep learning) (Marton & Säljö, 1976).

Causal Inference Expressions.

To identify causal reasoning, each message was coded for the presence of causal markers such as "because", "therefore", "so that", and "since". This approach aligns with work in computational linguistics and science education that treats causal connectives as indicators of mechanism-focused reasoning. Messages were labeled True or False depending on whether at least one causal marker appeared. In this educational context, causal markers serve as a proxy for mechanistic reasoning (Bachtar, Meulenbroeks, & van Joolingen, 2022). Their presence reflects a shift from descriptive questions toward explanatory inquiry, which characterizes progression in informal science learning.

Readability (Flesch–Kincaid Score).

Readability was computed using the Flesch-Kincaid Grade Level that estimates the U.S. school grade required to comprehend a text (Flesch, 1948). The score is based on sentence length and syllable count, with higher values indicating more complex language. This measure allowed comparisons between the linguistic complexity of user queries and AI responses. This metric

is leveraged to estimate the cognitive load imposed by the AI. A large disparity between user and AI readability scores may indicate that the system's expert register limits the learner's ability to process and extend the information.

Certainty Level.

Certainty was measured using a lexicon-based method adapted from Pei and Jurgens' (2021) framework for identifying expressions of confidence and uncertainty in natural language. Their approach models epistemic stance through lexical cues associated with strong commitment (e.g., definitely, certainly) and hedging or uncertainty (e.g., possibly, likely). Guided by this framework, each message received a certainty score reflecting the density of certainty-related expressions, with higher values indicating greater epistemic confidence. Certainty was computed separately for user queries and chatbot responses to examine differences in communicative stance between the two parties. Certainty in this study was treated as an indicator of epistemic stance (Rubin, 2010, Schiefer et al., 2022). High user certainty may reflect confirmation-oriented inquiry rather than exploratory reasoning, whereas AI certainty indexes the level of authority expressed in system responses. Disparities between the two can identify cases in which the system does not exhibit the degree of hedging expected in scientific discussion of complex environmental topics.

Results

What are in the user questions?

Users submitted brief questions ($M = 10.90$ words, $SD = 23.95$) that were low in Flesch-Kincaid scores (i.e., linguistic complexity) ($M = 6.35$, $SD = 4.81$) and expressed high levels of certainty ($M = 4.94$, $SD = 0.14$). In contrast, chatbot responses were substantially longer ($M = 102.53$ words, $SD = 76.77$), displayed higher linguistic complexity ($M = 13.91$, $SD = 5.73$), and conveyed lower certainty ($M = 4.63$, $SD = 0.53$). Question types in user-bot messages are presented in Figure 1. The majority of user questions were lower-order forms such as what, yes/no, or topic fragments. Approximately 3% of questions were why questions. There were also a few cases where users asked which-type questions that appeared to require more cognitive effort, indicating that they may have already formed a specific idea before seeking an answer. Expressions of causal reasoning in user messages were relatively uncommon, compared to the more frequent causal indications in the chatbot's language.

Across the dataset of 347 conversations, most interactions were short, typically containing fewer than five conversational turns. These indicate that users primarily engaged in surface-level information seeking rather than extended inquiry or multistep exploration.

How does the AI respond to user questions?

Figure 2 presents a view of the flow from types of questions users asked towards the reading level of Waterbot. A Kruskal-Wallis H test was conducted to examine differences in chatbot response complexity (Flesch-Kincaid Grade Level) across question types. The test indicated a significant effect of question type on response complexity: $H(9) = 213.89$, $p < .001$. Median Flesch-Kincaid scores differed by question type, with the highest medians for why (16.12), yes/no questions (15.47), and when (15.24). The lowest reading level from chatbot answers were for who questions (10.70). These results suggest that the chatbot generated more linguistically complex responses when addressing causal or explanatory prompts compared to factual or identification questions.

To examine alignment between chatbot and user language patterns, Wilcoxon signed-rank tests were conducted for three paired measures: readability, message length, and certainty. The test revealed significant differences for all three comparisons. In each round of conversation, chatbot messages were significantly more complex in readability measures than user messages ($W = 192,545.00$, $p < .001$). Chatbot responses were also longer ($W = 92,861.00$, $p < .001$) and expressed lower

certainty ($W = 764,598.00$, $p < .001$). These results indicate that, while the chatbot tended to produce longer and more syntactically complex messages than users, its language conveyed comparatively less confidence.

How do message characteristics relate to conversation depth?

To examine how session-level characteristics related to conversation length, Spearman's rank-order correlations were conducted between the total number of conversational turns and key linguistic or behavioral variables averaged across each session. Conversation length showed a strong positive correlation with chatbot word count ($\rho = .45$, $p < .001$), indicating that longer sessions tended to include more verbose chatbot responses. A smaller but significant positive association was also found with chatbot causal language ($\rho = .20$, $p < .001$) and user causal language ($\rho = .23$, $p < .001$). All other relationships, including those with readability (Flesch-Kincaid scores) and certainty measures, were nonsignificant ($ps > .05$). Overall, these results suggest that sessions with more turns were characterized by higher chatbot output and slightly greater use of causal expressions by both the user and the chatbot, but not by increased linguistic complexity or confidence.

Discussion

This pilot study examined how individuals used a text-based AI system to ask questions about water-related issues and how the system responded. By analyzing message-level features across conversational turns, the results provide an initial view of how users engage with an AI information tool in water as a scientific domain and how linguistic patterns shape the structure of these interactions. By analyzing message-level features across conversational turns, the results provide an empirical basis for refining the system to better support public engagement with complex science and public issues.

Across conversations, users submitted short, low-complexity questions that aligned with surface-level information seeking. Most questions were what or yes/no forms, and only a small proportion were why questions. These patterns are consistent with prior research showing that learners often begin with basic inquiries when entering complex scientific domains (Tawfik et al., 2020, Vattam et al., 2011). Causal language was uncommon in user messages, suggesting that most users approached the system with topic interest compared to more advanced conceptual frames. In contrast, the chatbot produced responses that were substantially longer and more linguistically complex. Response complexity varied by question type, with higher complexity for causal or explanatory prompts. Paired comparisons showed that the chatbot consistently used longer sentences, higher readability scores, and lower expressed certainty than users. This echoes prior studies that LLM explanations often mismatch intended education level (Joshi et al., 2025). Although these features reflect the system's tendency to provide thorough explanations, they also indicate a clear asymmetry between user inputs and system outputs. Conversation depth was generally limited, with most sessions ending after only a few turns. The only strong predictor of longer sessions was the chatbot's word count, and weaker associations were observed for causal language use. Taken together, these findings suggest that most users engaged with the system for quick informational purposes rather than extended inquiry.

From a learning-engineering perspective, these results highlight several alignment considerations for future Waterbot system design.

First, the finding that lengthy, complex responses correlate with shorter sessions suggested to us a modification of the system prompt. The redesigned Waterbot needs to prioritize brevity and linguistic simplicity in initial turns to match the user's input style, rather than providing exhaustive explanations by default. Second, the scarcity of causal (why) questions indicates a need for explicit scaffolding. To support deeper exploration, we could replace the passive interface with a suggestion engine that offers optional follow-up questions rooted in causal mechanisms. Finally, the analysis of certainty markers highlighted a need for better calibration of the AI's epistemic stance. Users may phrase questions as assertions or confirmations, or seek for

validation instead of instruction (Gall, 1985, Martin-Arbo, Castarlenas & Duenas 2021), but the AI hedging could be due to safety alignment or policy constraints (OpenAI, 2025). Therefore, the next design phase could incorporate a retrieval-augmented generation (RAG) framework that explicitly cites sources, allowing the system to express higher certainty on established facts while reserving hedging only for truly ambiguous policy areas.

As a pilot project, this study has several limitations. The dataset includes only text interactions without behavioral or contextual data. Certain technical factors, such as the method used to assign session identifiers, may also have influenced the count of unique conversations. Even so, the patterns identified here provide a useful starting point for future research. Later studies could introduce experimental variations in response style or assess user perceptions through real-time feedback collected in an updated system. In light of these, post-July 2025 versions of Waterbot include mechanisms to record real-time user feedback, improved logs of the information sources retrieved during responses, and an interface that supports easier follow-up questions. These updates make it possible to examine how users respond to explanations, what sources they find helpful, and whether lighter or more adaptive responses lead to deeper engagement. Overall, this pilot provides an initial baseline for understanding how people use conversational AI to learn about water-related topics. The findings contribute to ongoing efforts in learning engineering to refine AI systems through iterative analysis of learner traces.

Figure. 1

Distribution of User Inquiry Types for Waterbot (N = 2,978). The most popular questions were instructional prompts, followed by “what” questions

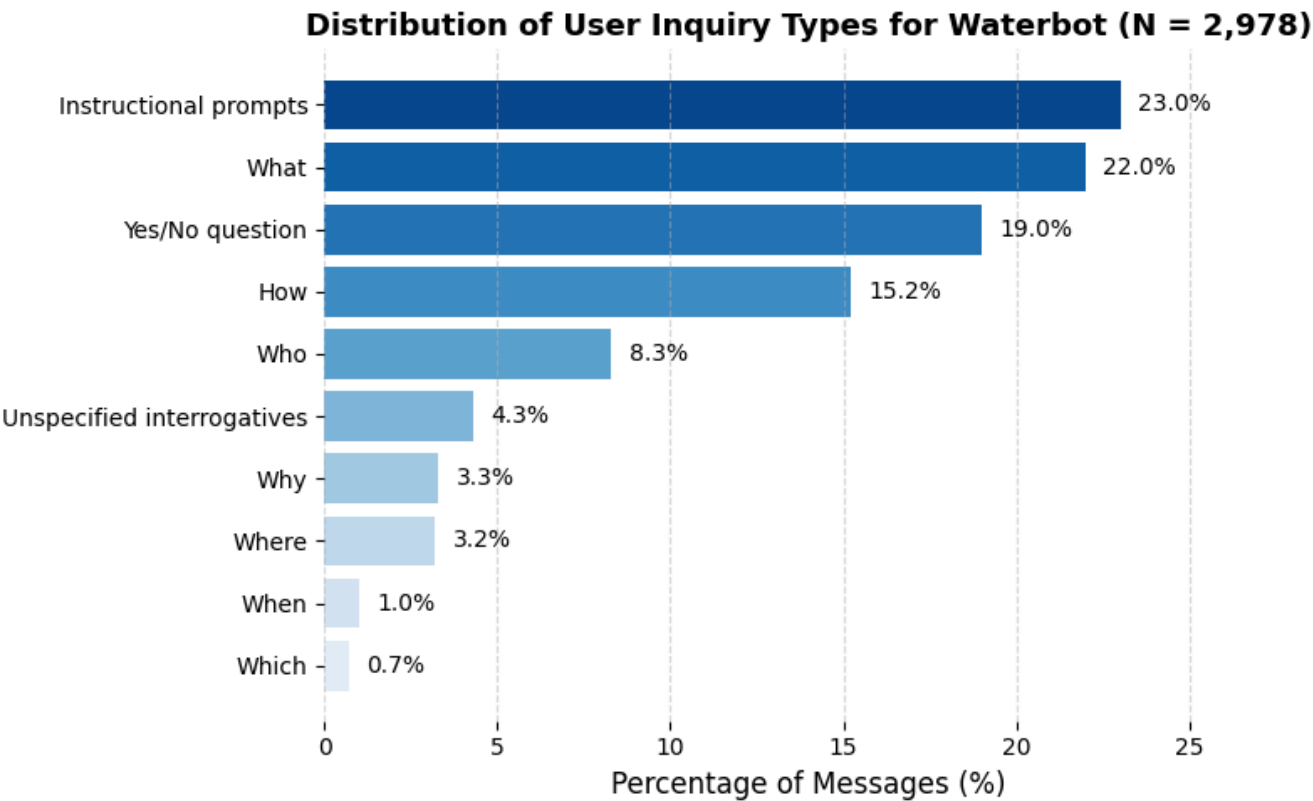
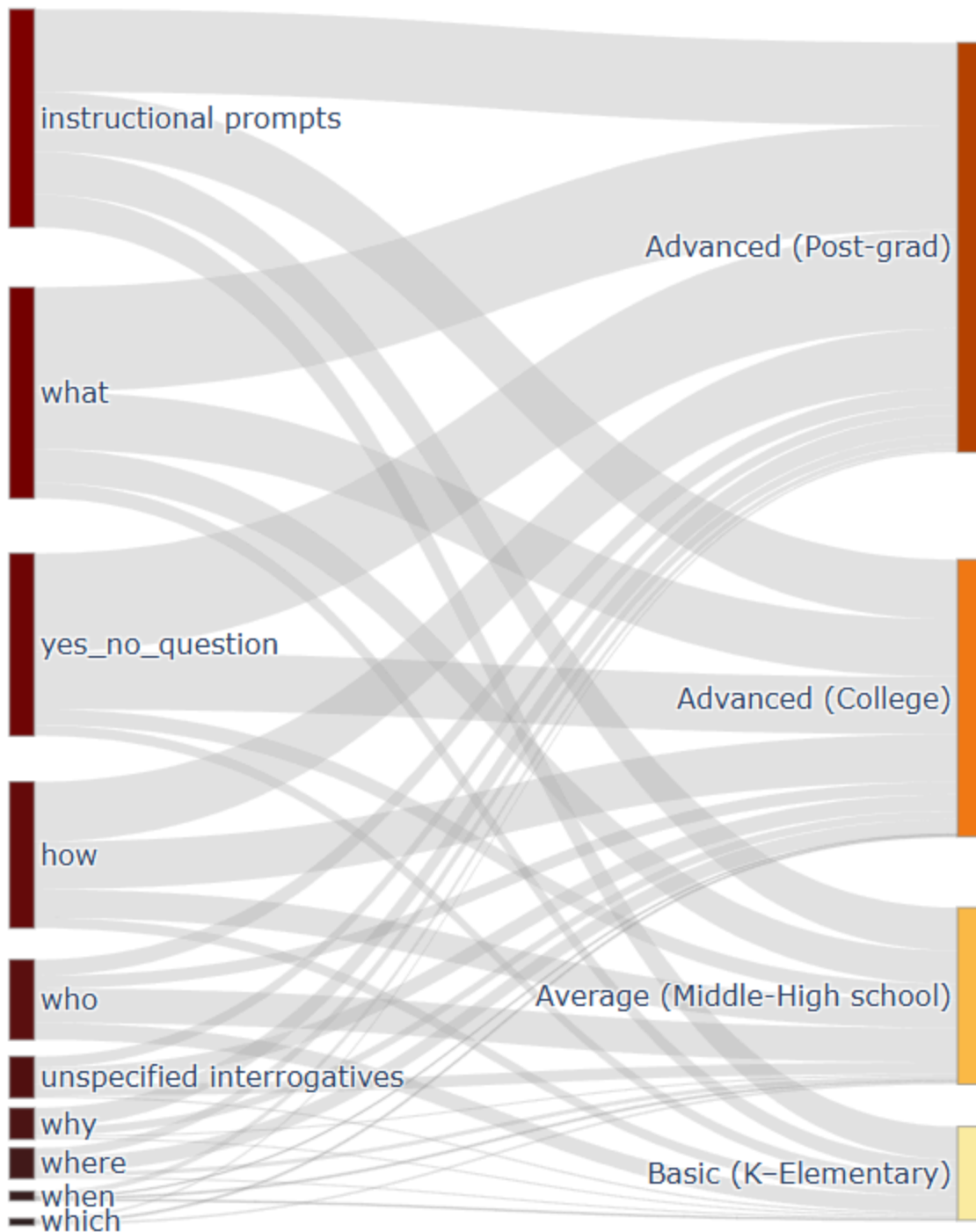


Figure. 2

Flow from User Question Types to Chatbot Readability Levels. Most chatbot responses are of advanced levels, and the answers are most complex for “why” questions.



Acknowledgments

The authors would like to acknowledge the generous support for this work provided by the Arizona Water Innovation Initiative (<https://azwaterinnovation.asu.edu/>).

References

- Bachtiar, R. W., Meulenbroeks, R. F., & van Joolingen, W. R. (2022). Mechanistic reasoning in science education: A literature review. *Eurasia Journal of Mathematics, Science and Technology Education*, 18(11), em2178.
- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed.
- Carradini, S. (2024). On the current moment in AI: Introduction to special issue on effects of artificial intelligence tools in technical communication pedagogy, practice, and research, part 1. *Journal of Business and Technical Communication*, 38(3), 187-198.
- Chin, C., & Osborne, J. (2008). Students' questions: a potential resource for teaching and learning science. *Studies in science education*, 44(1), 1-39.
- Craig, S. D., Avancha, K., Malhotra, P., Gorman, J. C., Verma, V., Likamwa, R., ... & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale*.
- Fleischer, H., Noglik, A., Borchers, C., & Schanze, S. (2025). Does Student Learning Rate Depend on Feedback Type and Prior Knowledge?.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233.
- Gall, S. N. L. (1985). Chapter 2: Help-seeking behavior in learning. *Review of research in education*, 12(1), 55-90.
- Goodell, J., Kolodner, J., & Kessler, A. (2022). Learning engineering applies the learning sciences. In *Learning engineering toolkit* (pp. 47-82). Routledge.
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. *Journal of Military Learning*, 7(1).
- IPSOS. (2025). Did you know? As more people engage with AI tools, concerns persist despite technology's recognized role in enabling progress. Ipsos. <https://www.ipsos.com/sites/default/files/ct/news/documents/2025-01/ipsos-essentials-infographic-january-2025.pdf>
- Joshi, B., He, K., Ramnath, S., Sabouri, S., Zhou, K., Chattopadhyay, S., Chattopadhyay, S., & Ren, X. (2025). ELI-Why: Evaluating the Pedagogical Utility of Language Model Explanations. *arXiv preprint arXiv:2506.14200*.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103, 102274.
- Kerry, A., Ellis, R., & Bull, S. (2008, December). Conversational agents in E-Learning. In *International conference on innovative techniques and applications of artificial intelligence* (pp. 169-182). London: Springer London.
- Martin-Arbo, S., Castarlenas, E., & Duenas, J. M. (2021). Help-seeking in an academic context: A systematic review. *Sustainability*, 13(8), 4460.
- Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British journal of educational psychology*, 46(1), 4-11.
- OpenAI. (2025, January 23). Operator system card. <https://openai.com/index/operator-system-card>

- Pei, J., & Jurgens, D. (2021, November). Measuring Sentence-Level and Aspect-Level (Un) certainty in Science Communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 9959-10011).
- Rajan, B., Carradini, S., & Lauer, C. (2024, May). The Arizona water chatbot: Helping residents navigate a water uncertain future one response at a time. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-10).
- Rubin, V. L. (2010). Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management*, 46(5), 533-540.
- Schiefer, J., Edelsbrunner, P. A., Bernholt, A., Kampa, N., & Nehring, A. (2022). Epistemic beliefs in science—a systematic integration of evidence from multiple studies. *Educational Psychology Review*, 34(3), 1541-1575.
- Su, J., Healey, J., Nakov, P., & Cardie, C. (2025). Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms. *arXiv preprint arXiv:2505.00127*.
- Tawfik, A. A., Graesser, A., Gatewood, J., & Gishbaugher, J. (2020). Role of questions in inquiry-based instruction: Towards a design taxonomy for question-asking and implications for design. *Educational Technology Research and Development*, 68(2), 653-678.
- Terzimehić, N., Bühler, B., & Kasneci, E. (2025). Conversational AI as a Catalyst for Informal Learning: An Empirical Large-Scale Study on LLM Use in Everyday Learning. *arXiv preprint arXiv:2506.11789*.
- Thai, K. P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. (2022). Learning engineering is human-centered. In *Learning engineering toolkit* (pp. 83-123). Routledge.
- Vattam, S. S., Goel, A. K., Rugaber, S., Hmelo-Silver, C. E., Jordan, R., Gray, S., & Sinha, S. (2011). Understanding complex natural systems by articulating structure-behavior-function models. *Journal of Educational Technology & Society*, 14(1), 66-81.
- Yona, G., Aharoni, R., & Geva, M. (2024). Can large language models faithfully express their intrinsic uncertainty in words?. *arXiv preprint arXiv:2405.16908*.



Design and Pilot Evaluation of a Gamified Narrative Chatbot for STEM Education

Maria Goldshtein, Ishrat Ahmed, Vipin Verma, Tracy Arner, & Danielle S. McNamara

Gamified learning

GenAI

Socratic method

STEM education

Extended Abstract

Introduction

The rise and ease of access to Generative Artificial Intelligence (GenAI) opens up opportunities for personalized experiences in many spheres, including education. Research has shown that learning can be supported through gamification (Plass et al., 2015) by enhancing motivation and engagement, Socratic instruction (Favero et al., 2024) which fosters critical thinking through guided questioning and by narrative storylines (Gee, 2003) which promote immersion and meaning making in educational contexts. The current presentation showcases the nested learning engineering cycle (Goodall & Kolodner, 2023; Totino & Kessler, 2024), including the design process, a pilot study, and the preliminary results for three gamified chatbots incorporating GenAI and narrative writing to help students learn a STEM topic.

Method and Analysis

Three chatbots incorporating GenAI were designed in an iterative prompt engineering and testing process, utilizing ASU's CreateAI Builder platform (Ahmed et al., 2025). The goal of these chatbots was to teach participants about Photosynthesis through a socratic method. Two of the three chatbots were gamified (reward-based and role-based) and included a narrative storyline. The third was a baseline chatbot without gamified or narrative features designed for comparing results.

We conducted a pilot study to test these chatbots with 53 undergraduate student participants through ASU's SONA psychology subject pool. Participants filled out a pre-study quiz about the topic taught by the chatbot, interacted with one of the three chatbots to solve 3-4 challenges, then filled out a post-interaction knowledge quiz, followed by a survey related to the user experience with the chatbot they interacted with.

Pre- and post-test scores were compared for each participant. We also analyzed chatbot interaction logs for bot response quality and participant engagement. The surveys were qualitatively and quantitatively analyzed to test participant learning, engagement, and participants' prior experience with games, text-based games and GenAI tools.

Findings

Preliminary findings show that participants found the chatbot interactions to be engaging, coherent, and contributing to their knowledge on the topic taught. Despite all chatbot versions contributing to student learning, participants reported enjoying the gamified features offered in both gamified chatbots tested. We will report specific findings about the chatbot features and user experiences with the chatbots and other similar technologies.

Implications. Study findings contribute to the understanding of the novel application of GenAI for gamified learning environments and the interaction of GenAI with socratic learning support. Findings will inform future iterations of the chatbot, with the goal of being employed as a tool in classes, or a supplementary tool for learners of STEM topics.

Acknowledgments

We would like to acknowledge work done by Alison Nguyn in supporting parts of the analysis. This work has been supported by

References

- Ahmed, I., Alvarado, P., Jain, S., Arner, T., Reilley, E., & McNamara, D. S. (2025). Arizona State University CreateAI Platform. Grantee Submission.
- Favero, L., Pérez-Ortiz, J. A., Käser, T., & Oliver, N. (2024, October). Enhancing critical thinking in education by means of a Socratic chatbot. In *International Workshop on AI in Education and Educational Research* (pp. 17-32). Cham: Springer Nature Switzerland.
- Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in entertainment (CIE)*, 1(1), 20-20. <https://doi.org/10.1145/950566.950595>
- Goodell, J., & Kolodner, J. (Eds.). (2023). *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond*. Taylor & Francis.
- Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, 50(4), 258-283. <https://doi.org/10.1080/00461520.2015.1122533>

Totino, L., & Kessler, A. (2024). "Why did we do that?" A Systematic Approach to Tracking Decisions in the Design and Iteration of Learning Experiences. *The Journal of Applied Instructional Design*, 13(2) <https://doi.org/10.59668/1269.15630>



Designing for Student Engagement with AI in Courseware: Lessons from Iterative Improvements to DOT in REAL CHEM

Abigail Stein, Shang-Ting Ciou, Tingyue Cui, Kimberly Larson, Lilly Lee, Sherry Li, Chris Mead, Ashley Xu, & David J. Yaron

AI

Courseware

learning engineering

Student Engagement

This design-based research study reports on a Learning Engineering (LE) cycle embedded within the broader REAL CHEM courseware initiative, examining how students use and perceive DOT, a generative AI tutor for general chemistry. In Fall 2024, student awareness and engagement with DOT were low, and interviews revealed reliance on external AI tools and limited generative use. Guided by LE principles—human-centered design, iterative refinement, and evidence-informed decision making—we redesigned DOT by refining its base prompt and introducing AI Activation Points to make the system more proactive and contextually aligned with student workflows. Summer 2025 findings showed increased awareness, higher engagement, and improved student satisfaction. Results illustrate how targeted, iterative LE cycles can improve the effectiveness of AI-enabled learning supports at scale.

Introduction

REAL CHEM is a fully instrumented courseware environment for general chemistry developed at Carnegie Mellon University and Arizona State University (CMU, 2022) and built around the OpenStax Chemistry 2e textbook (Flowers et al., 2009). REAL CHEM integrates content, assessment, and analytics into a coherent instructional system designed to support student learning, particularly for students historically underserved in STEM.

The design and evolution of REAL CHEM reflect core commitments of Learning Engineering (LE), an interdisciplinary field that integrates learning sciences, instructional design, data, and human-centered methods to iteratively improve learning systems through evidence (Baker et al., 2022; Goodell et al., 2023). Within this larger effort, this paper focuses on a bounded LE cycle centered on DOT, the Digital Online Tutor.

DOT is a generative AI tutor embedded within the REAL CHEM interface. At the outset of this study, DOT functioned as an omnipresent chat tool available to students but rarely used. This paper documents how learner data and interviews were used to diagnose challenges related to awareness, trust, and use; inform targeted design decisions; and evaluate the impact of those decisions.

Methods

Our methods follow a design-based research approach (Cobb et al., 2003) aligned with LE's iterative design-test-refine cycles. In Fall 2024, we conducted semi-structured interviews with 10 undergraduate students enrolled in REAL CHEM-supported

courses and analyzed DOT interaction logs from 473 students. Interview questions focused on awareness of DOT, use of AI tools (e.g., ChatGPT), trust and accuracy perceptions, and decisions about when to seek AI support.

Findings from this cycle informed targeted design updates to DOT. In Summer 2025, we conducted a second round of interviews (n=6) and analyzed interaction logs (n=172 students). Interview protocols were extended to probe student experiences with the redesigned DOT and AI Activation Points.

Student-DOT interactions were coded into categories (e.g., content, course administration, platform issues, feedback, and unserious conversation). Descriptive statistics were used to examine engagement patterns across cycles.

Results

In Fall 2024, student awareness and use of DOT was extremely limited. Half of interviewed students were unaware of the feature, and only 11% of the 4,249 enrolled students interacted with DOT, most fewer than five times. Students reported relying on external AI tools (e.g., ChatGPT, Gemini, Claude) and traditional resources such as peers, instructors, and YouTube. Interaction data showed that queries were dominated by copied course questions (32%) and factual recall (29%), with limited generative or explanatory use. Students also cited concerns about accuracy, verbosity, and poorly rendered mathematical notation.

Based on these findings, we made two design changes. First, we refined DOT's base prompt to prioritize concise, stepwise explanations, avoid directly providing final answers, and reduce mathematical and formatting errors. Second, we introduced AI Activation Points—structured moments where DOT proactively engages students at key points in the courseware—to improve visibility and alignment with student workflows.

Summer 2025 findings showed substantial improvement. All interviewed students were aware of DOT, overall interaction rates increased to 36.2%, and student satisfaction ratings averaged 4.5/5. Activity-level and open-ended Activation Points were perceived as timely and helpful, while page-level activations showed low engagement, suggesting the importance of precise timing. Copying and pasting course questions into DOT decreased.

Discussion

These findings demonstrate how a focused LE cycle can improve student engagement with AI tutors embedded in courseware. Rather than treating DOT as a static feature, learner data and qualitative insights were used to guide targeted, human-centered design decisions and evaluate their impact. Results help explain why students often prefer general-purpose AI tools over course-specific tutors and show how intentional design—proactivity, contextual relevance, and workflow alignment—can narrow that gap.

More broadly, this study illustrates the value of nested LE cycles within large-scale courseware initiatives, enabling systematic improvement of AI-enabled supports while remaining responsive to student needs.

Acknowledgments

We gratefully acknowledge the financial support from The Bill & Melinda Gates Foundation to this research effort. Opinions and conclusions expressed in this article do not necessarily reflect the views of this funding agency.

References

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*. <https://doi.org/10.1037/tmb0000058>
- Carnegie Mellon University. (2022, June 8). Universities partner to make chemistry more equitable. <https://www.cmu.edu/news/stories/archives/2022/june/asu-cmu-chemistry.html>
- Cobb, P., Confrey, J., DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational researcher*, 32(1), 9-13.
- Flowers, P., Theopold, K., Langley, R., & Robinson, W. (2019). Chemistry 2e. OpenStax. <https://openstax.org/details/books/chemistry-2e>
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning Engineering at a Glance. *Journal of Military Learning*. <https://www.armyupress.army.mil/Portals/7/journal-of-military-learning/images/Conference-Edition-2023/JML-Conference-Edition-2023-TOC-v1.pdf>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Guggemos, J., Opfermann, M., Schmid, S., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Instruction*, 88, 101666.



EdLight Research Portal: An Expert-Annotated Repository of Handwritten Math Student Work

Michel Meneses & Selena Castro

formative assessments

Handwritten math dataset

K-12 student reasoning

This paper introduces the EdLight Research Portal (ERP), a publicly available repository of expert annotated handwritten math student work. ERP addresses the gap between formative assessment practices in real mathematics classrooms and the digital datasets commonly used in AI-powered education research. The dataset captures authentic paper-based reasoning from 5th to 9th grade students and includes standardized rubric scores, misconception tags, and contextual metadata generated by trained math educators. ERP enables researchers to study student thinking at scale, examine patterns in mathematical misconceptions, and benchmark machine learning models on tasks that reflect real classroom data. The portal provides a foundation for advancing transparent, pedagogically grounded research in AI applications for mathematics education.

Introduction

Learning Engineering (LE) integrates learning sciences with engineering rigor to iteratively improve educational systems through data-informed practice (Baker et al., 2022). In K-12 mathematics, LE can alleviate teacher burdens, such as limited instructional time, by automating routine assessments and refining recommendation systems (Kadaruddin, 2023; Noroozi et al., 2024). However, rapid LE research is currently hindered by a lack of high-quality, annotated datasets that reflect authentic classroom conditions. While existing repositories often rely on typed or digital ink inputs (Gervais et al., 2024; Zhang et al., 2024), they fail to capture the nuanced reasoning found in traditional handwritten work, the primary mode of formative assessment in schools.

To bridge this gap, we introduce the EdLight Research Portal (ERP): a public repository of 19K expert-annotated samples of authentic handwritten K-12 mathematics. Developed as a nested LE cycle within the broader EdLight platform, each sample features rubric scores, misconception tags, and metadata produced by trained educators. By providing this domain-specific infrastructure, we enable researchers to study student thinking at scale, identify mathematical misconception patterns, and benchmark machine learning models against data that truly reflects the classroom experience.

Methods

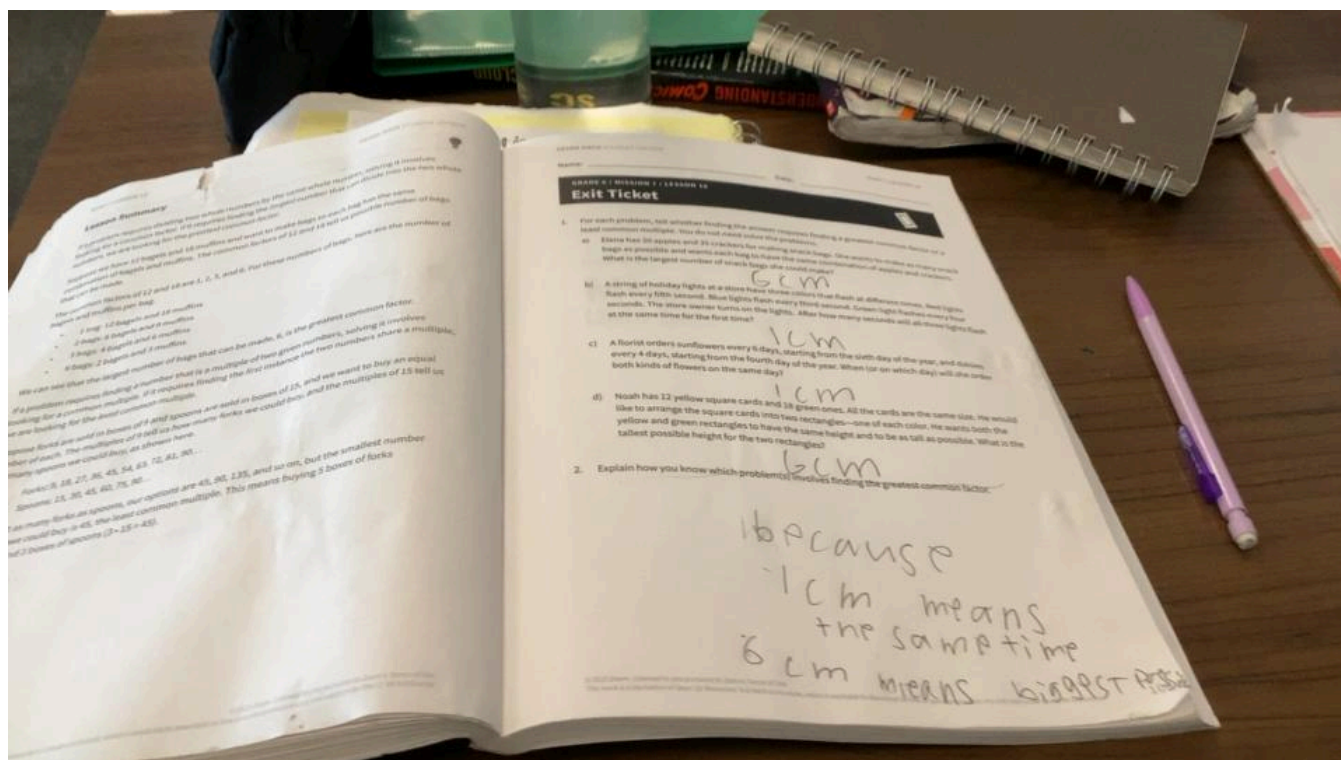
Data Collection and Preparation

The dataset was collected in collaboration with three EdLight partner schools, where students completed paper-and-pencil mathematics assignments during regular classroom instruction. These artifacts were scanned or photographed by teachers using EdLight's platform and securely uploaded to a centralized repository. Each submission was paired with metadata provided by teachers, including grade level, assignment details, learning standards, and anonymized student demographics when available. All student work samples as of November 2025 come from 5th to 9th grade Illustrative Mathematics exit tickets, known as Cool Downs.

To ensure high-quality annotations at scale, EdLight hired a team of 55 experienced K-12 math teachers, referred to as math instructional specialists (MIS), who conducted all labeling activities between November 2023 and June 2024. These specialists were trained to assign misconception tags and overall rubric scores based on standardized scoring guidelines aligned with state and national math standards. Their work produced expert-labeled samples that preserve the diversity and authenticity of real student reasoning while maintaining consistency across thousands of annotations.

Figure 1.

A student-work image sample from the released dataset. It corresponds to an Illustrative Mathematics exit ticket (i.e., "Cool Down") for Lesson 18 of Unit 7 in 6th grade.



Annotation of Student Work

Each student response was evaluated along two dimensions: overall rubric score and mathematical misconceptions. Rubric scores range from 0 to 3, where 3 indicates a complete and correct solution, 2 reflects conceptual understanding with minor errors or incomplete reasoning, 1 indicates significant errors, and 0 shows no viable mathematical strategy.

Misconception annotations were developed in collaboration with district math leaders and one of the authors of the Illustrative Mathematics middle school curriculum. This collaboration ensured that the schema aligns with instructional expertise, reflects curricular intent, and captures the recurring reasoning and procedural errors students commonly make. Responses could receive multiple tags, including repeated instances of the same category, enabling granular characterization of student thinking. Responses without observable errors were tagged "N/A – No misconceptions present." This dual-layer approach captures both correctness and reasoning quality, supporting analyses beyond binary right-or-wrong evaluation.

Table 1.
Misconception tags developed by EdLight.

Misconception	Definition
Computation error	Mistakes in arithmetic operations such as addition, subtraction, multiplication, or division.
Precision error	Mistakes related to numerical accuracy or formatting.
Conceptual misunderstanding	A lack of understanding or incorrect interpretation of mathematical concepts or principles.

Representational error	Mistakes in expressing or interpreting mathematical information through visual representations.
Insufficient explanation	Inadequate explanation or model.
Incomplete	Students didn't finish or didn't answer the question at all.
Did not follow	Refers to cases where students did not follow directions.
Unable to diagnose	There is insufficient visible evidence to determine any other type of misconception.

Quality Assurance and Scoring Reliability

A structured quality assurance process was implemented to ensure consistency in both overall rubric scoring and misconception tagging. During the initial phase of annotation, team members manually conducted inter-rater reliability checks to align interpretations of the scoring criteria and misconception definitions. This calibration process included independent scoring of shared samples and group discussions to resolve discrepancies and refine annotation guidelines.

Following the transition to a web-based data annotation platform, quality assurance procedures were maintained through systematic cross-annotation and expert auditing. A subset of student responses was assigned to multiple annotators to monitor agreement across scorers. Senior instructional specialists conducted weekly audits, reviewing approximately ten percent of all completed annotations to verify adherence to scoring standards and misconception criteria. Feedback from these audits was shared regularly with annotators and informed ongoing refinements to training materials and annotation procedures.

Results

The EdLight Research Portal (ERP) is publicly available at <https://research-portal.edlight.com> (access can be requested via the form at <https://api.edlight.com/contact/>). ERP contains 19,641 samples of handwritten mathematics student work within 5th to 9th grades. Each sample includes a link to the corresponding image, metadata such as assignment name, content standard, grade level, and available demographic information including student gender and ethnicity. In addition, all samples are annotated with expert-generated misconception tags and overall rubric scores. Table 2 summarizes the distribution of samples across grade levels and mathematical content areas.

Table 2.
Count and percentage of ERP's samples across content areas by grade.

Content area	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9
Expressions & equations	198 (30%)	17 (3%)	4,446 (33%)	1,486 (31%)	166 (87%)
Functions	-	-	-	158 (3%)	15 (8%)
Geometry	48	-	1,030	1,462	8

	(7%)		(8%)	(31%)	(4%)
Number System	79 (12%)	373 (66%)	5,014 (37%)	642 (13%)	2 (1%)
Ratios & Proportions	342 (51%)	173 (31%)	2,665 (20%)	726 (15%)	-
Statistics & Probabilities	-	-	284 (2%)	307 (6%)	-
Total student-assignments	667 (100%)	563 (100%)	13,439 (100%)	4,781 (100%)	191 (100%)

Table 3 presents the frequency and proportion of misconception tags assigned to student responses. Because a single sample may contain more than one type of misunderstanding, multiple tags can be applied to the same piece of work. Figure 2 complements this analysis by showing the distribution of overall rubric scores across content areas, allowing for comparisons between student performance and the prevalence of conceptual challenges.

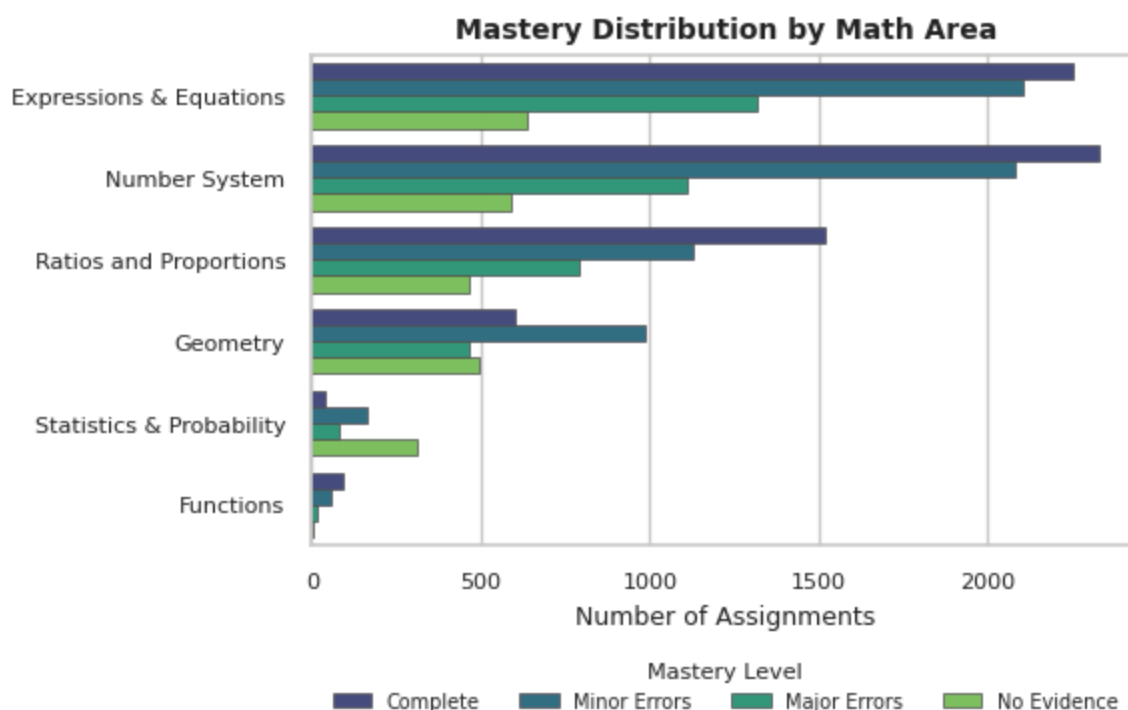
Table 3.

Frequency and proportion of misconception tags assigned to student responses in the EdLight Research Portal. Multiple tags can be applied to a single sample, reflecting the presence of multiple reasoning errors within individual student work.

Misconception	Percent of student work
Computation error	11%
Precision error	4%
Conceptual misunderstanding	38%
Representational error	4%
Insufficient explanation	12%
Incomplete	6%
Did not answer specific question	2%
Unable to diagnose	11%
No misconception present	36%

Figure 2.

Distribution of overall rubric scores across content areas in the EdLight Research Portal, illustrating variations in student performance and enabling comparison with the prevalence of annotated misconceptions. Mastery level descriptors are abbreviated for legibility.



Discussion

The EdLight Research Portal provides foundational infrastructure for Learning Engineering research in K–12 mathematics by capturing authentic paper based student work at scale. By preserving handwritten reasoning together with rubric aligned evaluations, misconception tags, and classroom relevant metadata, ERP reflects formative assessment as it occurs in everyday instructional practice. This fidelity to classroom conditions addresses a central challenge in Learning Engineering by enabling data informed, iterative improvement of educational systems grounded in real student thinking.

As a public, educator annotated resource embedded within an ongoing Learning Engineering cycle, ERP supports the study of learning progressions, misconception patterns, and instructional decision making in authentic contexts. It also enables the development and evaluation of computational tools, including AI assisted assessment, that are designed to reduce teacher burden while remaining aligned with instructional intent. Future work will expand the dataset with additional solution level annotations and examine how such tools can be integrated into formative assessment workflows to support scalable improvement in mathematics education.

Acknowledgments

We thank the Gates Foundation for supporting this work, the Research Partnership for Professional Learning (RPPL) and Annenberg Institute at Brown University for analyses informing the dataset design, HackSoft for engineering support, and the educators and district partners whose expertise made this dataset possible.

References

Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*, 3(1), 1–23. <https://doi.org/10.1037/tmb0000058>

Gervais, P., Fadeeva, A., & Maksai, A. (2024). MathWriting: A Dataset For Handwritten Mathematical Expression Recognition. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*.

Kadaruddin, K. (2023). Empowering Education through Generative AI: Innovative Instructional Strategies for Tomorrow's Learners. *International Journal of Business, Law, and Education*.

Noroozi, Omid et al. (2024). Generative AI in Education: Pedagogical, Theoretical, and Methodological Perspectives. *International Journal of Technology in Education*.

Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K., Gao, P., & Li, H. (2024). MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? *European Conference on Computer Vision*.



Epistemic Cognition and Uncertainty Navigation with a Domain-Specific AI Chatbot in STEM Education

Yiwen Li, Chengshuai Zhao, Garima Agrawal, Yuli Deng, Jongchan Park, Ying-Chih Chen, & Huan Liu

AI-assisted learning

Computer Science Education

Student Uncertainty

Domain-specific AI chatbots are increasingly embedded in STEM courses, yet little is known about how students judge, interpret, and act on AI-generated feedback during learning. This study examines students' epistemic cognition when interacting with Cyberbot, a domain-specific AI-powered chatbot integrated into a cloud computing course. Drawing on post-course semi-structured interviews with twelve students, we analyze how learners articulate epistemic aims, evaluate the correctness and reliability of AI feedback, and navigate uncertainty when deciding whether to revise or persist in their problem-solving approaches. Findings indicate that Cyberbot does not eliminate uncertainty for learners; instead, it reorganizes epistemic work by shifting responsibility toward verification, judgment, and decision-making. Students demonstrated calibrated trust, actively cross-checking AI responses with other sources and using the chatbot primarily for epistemic verification rather than open-ended exploration. These results highlight the importance of examining epistemic processes when evaluating AI-supported learning.

Introduction

AI-powered chatbots are increasingly embedded in STEM education to provide explanations, feedback, and on-demand support during problem-solving activities. In response to concerns about the limitations of generic large language models, recent research has emphasized the pedagogical potential of domain-specific AI chatbots that are explicitly aligned with curricular goals and disciplinary content. Empirical studies of such systems have reported positive effects on learning performance, situational interest, and student acceptance in STEM contexts, particularly when chatbots are customized to specific subject domains (Rücker & Becker-Genschow, 2025).

These questions are especially salient in computing-related STEM domains such as cloud computing, where problem solving often involves ambiguous requirements, evolving system constraints, and multiple acceptable solutions. In such contexts, students frequently encounter uncertainty arising from incomplete understanding, conflicting information, or doubts about whether their solutions adequately meet task demands. Rather than viewing uncertainty as a barrier to learning, research has emphasized its role in prompting inquiry, reflection, and sensemaking (Manz, 2015; Berland et al., 2016).

To address this gap, the present study adopts an epistemic cognition perspective to examine how students engage with a domain-specific AI chatbot during moments of uncertainty. Epistemic cognition concerns how learners think about knowledge, evidence, and justification, including how they evaluate the correctness and trustworthiness of information and decide whether

to revise or persist in their approaches (Chinn et al., 2011; Greene et al., 2014). This study investigates students' epistemic cognition during interactions with Cyberbot, a domain-specific AI-powered chatbot embedded in a cloud computing course. Drawing on post-course interviews, we examine how students evaluate AI feedback, navigate uncertainty, and position the chatbot within their broader epistemic ecosystem of peers, instructors, and other resources.

Conceptual Framework

Epistemic cognition refers to how individuals think about knowledge, evidence, and justification, including their epistemic aims, ideals, and processes (Chinn et al., 2011). In learning contexts, epistemic cognition shapes how students determine what counts as a satisfactory explanation, how they evaluate information sources, and how they justify decisions during problem solving. Prior research has emphasized that epistemic cognition is situated and task-dependent, varying across domains and learning environments (Greene et al., 2014).

Uncertainty is a central condition under which epistemic cognition becomes visible. In STEM learning, uncertainty often arises from incomplete knowledge, ambiguous information, or competing explanations, prompting learners to engage in epistemic judgment and sensemaking (Chen et al., 2024). Rather than conceptualizing uncertainty as a negative outcome to be minimized, this study treats uncertainty as a productive condition that activates epistemic processes, such as evaluating evidence, seeking verification, and deciding whether to revise or persist in an approach (Berland et al., 2016).

In AI-supported learning contexts, epistemic cognition is particularly salient because students must determine how to interpret and use AI-generated feedback. Interacting with AI systems requires learners to judge the reliability, sufficiency, and relevance of responses, especially when feedback is incomplete or conflicts with prior knowledge. This framework allows us to examine how students enact epistemic cognition while navigating uncertainty during interactions with a domain-specific AI chatbot.

Context and Methods

This study was conducted in a semester-long cloud computing course at a large public university. As part of the course, students used Cyberbot, a domain-specific AI-powered chatbot designed to provide course-aligned explanations and clarification during project work and exam preparation. Unlike open-domain chatbots, Cyberbot was tailored to the course content and instructional goals, with the intent of supporting learning rather than providing direct solutions.

Post-course semi-structured interviews were conducted with twelve students. Interview questions focused on students' epistemic aims when using Cyberbot, their expectations for AI feedback, how they evaluated the correctness and reliability of responses, and how they made decisions under uncertainty. Interview data were analyzed thematically to identify recurring patterns in students' epistemic cognition and uncertainty navigation during interactions with Cyberbot.

Findings

Across interviews, students described using Cyberbot primarily as a tool for epistemic verification rather than open-ended exploration or solution generation. When encountering uncertainty—such as unclear task requirements or doubts about correctness—students turned to the chatbot to assess whether their current approach aligned with course expectations. Importantly, students did not treat Cyberbot as an authoritative source; instead, they demonstrated calibrated trust by actively evaluating the correctness and reliability of AI feedback based on consistency with lectures, course materials, and prior knowledge. When responses were vague or conflicted with other sources, students interpreted these moments as signals to seek verification elsewhere rather than accepting AI feedback uncritically. These practices indicate that interacting with Cyberbot required ongoing epistemic judgment.

Cyberbot also shaped how students navigated uncertainty when deciding whether to revise or persist in their problem-solving approaches. Rather than prompting wholesale strategy changes, AI feedback most often reinforced confidence in existing plans or supported incremental adjustments. Students weighed AI suggestions against task constraints and their own reasoning, highlighting uncertainty navigation as a deliberative epistemic process. At the same time, students consistently positioned Cyberbot as one epistemic resource within a broader learning ecosystem, integrating it with peers, instructors, and documentation. This positioning underscores that learning with a domain-specific AI chatbot involved distributed epistemic work, in which judgment and verification were coordinated across multiple sources rather than delegated to the AI alone.

Discussion

These findings suggest that domain-specific AI chatbots do not eliminate uncertainty for learners but instead reconfigure epistemic responsibility by foregrounding judgment, verification, and decision-making. Rather than passively accepting AI-generated feedback, students engaged in calibrated evaluation, using Cyberbot to assess alignment with task expectations while retaining responsibility for determining correctness and adequacy. Uncertainty functioned as a productive condition that prompted epistemic action, shaping when students chose to revise or persist and how they integrated AI feedback with other sources. Viewed through an epistemic cognition lens, AI-supported learning in this context involved distributed sensemaking across human and non-human resources, highlighting the importance of designing AI tools that support learners' epistemic judgment rather than providing definitive answers.

Acknowledgments

This research was supported by the National Science Foundation under Grant No. [2335666] and [2404966]. Any opinions, findings, and conclusions or recommendations expressed in this research are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- Berland, L. K., Schwarz, C. V., Krist, C., Kenyon, L., Lo, A. S., & Reiser, B. J. (2016). Epistemologies in practice: Making scientific practices meaningful for students. *Journal of Research in Science Teaching*, 53(7), 1082-1112.
- Chen, Y. C., Jordan, M., Park, J., & Starrett, E. (2024). Navigating student uncertainty for productive struggle: Establishing the importance for and distinguishing types, sources, and desirability of scientific uncertainties. *Science Education*, 108(4), 1099-1133.
- Chinn, C. A., Buckland, L. A., & Samarapungavan, A. L. A. (2011). Expanding the dimensions of epistemic cognition: Arguments from philosophy and psychology. *Educational psychologist*, 46(3), 141-167.
- Greene, J. A., & Seung, B. Y. (2014). Modeling and measuring epistemic cognition: A qualitative re-investigation. *Contemporary Educational Psychology*, 39(1), 12-28.
- Manz, E. (2015). Representing student argumentation as functionally emergent from scientific activity. *Review of Educational Research*, 85(4), 553-590.
- Rücker, C. R., & Becker-Genschow, S. (2025). Enhancing Enthusiasm for STEM Education with AI: Domain-Specific Chatbot as Personalized Learning Assistant. *Computers and Education Open*, 100315.



From Course Concept to Lecture Video: An AI-Powered System for Automated MOOC Development

Jacob Igo, Huaiyuan Yao, Wanpeng Xu, Nadia Kellam, & Hua Wei

AI-generated video

ICAP framework

MOOC development

Producing lecture videos for online courses requires significant time and expertise, and current AI-generated videos often lack accuracy, consistency, and pedagogical grounding. This study presents a controllable system that converts slides and scripts into lecture videos. The system follows instructor intent and integrates the ICAP framework to ensure the presence of constructive questions. We evaluate the system using four metrics: ICAP assurance, video generation time ratio, speech consistency, and slides coverage rate. Results show that the system provides higher controllability and stronger educational alignment than NotebookLM, demonstrating its potential for scalable MOOC development.

Introduction

Massive Open Online Courses continue to expand access to higher education, yet instructors face three major challenges when producing lecture videos. First, video creation is highly time-consuming and requires substantial effort in design and production, which slows course development and affects the quality of online materials (Hollands and Tirthali, 2014; Guo et al., 2014). Second, current video generation AI tools like NotebookLM (Google, 2025) often produce inconsistent or inaccurate content and cannot reliably follow instructor-provided slides or scripts. Lastly, existing systems do not incorporate learning science principles, which reduces opportunities for meaningful cognitive engagement (Bauer, 2025).

This study presents an automated system that converts course slides and scripts into lecture videos through a controllable generation process to address these challenges. The system ensures accurate alignment with instructor intent, maintains full coverage of the instructional materials, and integrates the ICAP framework (Chi and Wylie, 2014), which characterizes different levels of cognitive engagement: Interactive, Constructive, Active, and Passive. By embedding constructive and interactive questions, the system encourages learners to move beyond passive reception and engage in deeper processing of instructional content.

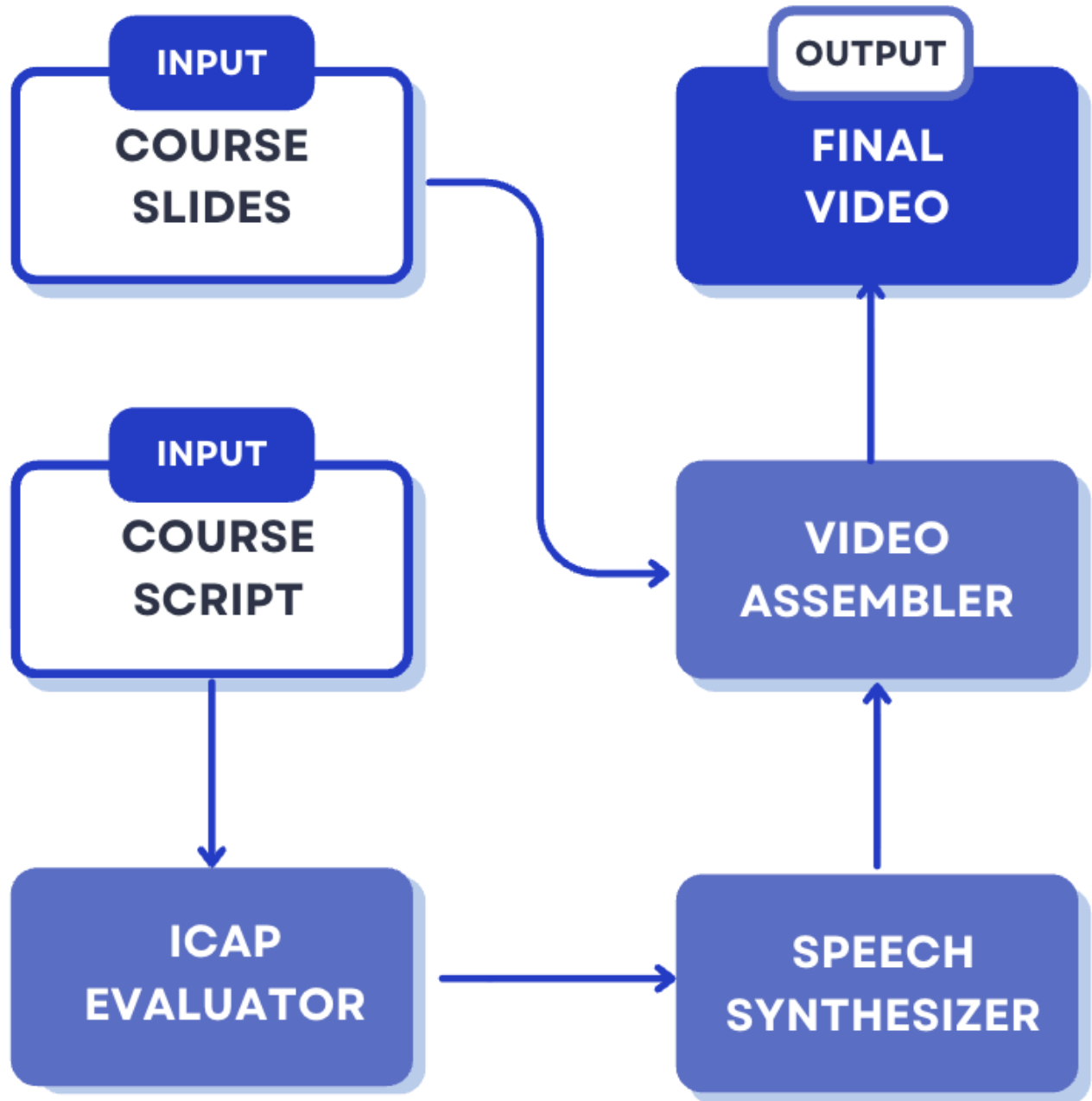
Our contributions are threefold. First, we propose a controllable pipeline for automated MOOC video generation. Second, we integrate ICAP-based design principles to promote higher levels of learner engagement. Finally, we evaluate the system using metrics that directly address current challenges in consistency, efficiency, and pedagogical quality.

Methodology

The proposed system automatically generates MOOC lecture videos by transforming course materials into coherent presentations. It comprises three sequential modules ICAP Evaluator, Speech Synthesizer, and Video Assembler that ensure pedagogical alignment, generation efficiency, and content consistency under the ICAP framework, as shown in Figure 1.

Figure. 1.

Overview of the proposed pipeline. Course slides and scripts are used as inputs, with the script evaluated by the ICAP module before speech synthesis and video assembly produce the final lecture video.



The system takes course slides and accompanying scripts as input. Instructors may upload pre-existing materials, or, when only a course title or outline is provided, Instructional Agents (Yao et al., 2025) automatically generate slides and scripts aligned with the specified learning objectives using large language models. Subsequently, the ICAP Evaluator employs a large language model to analyze the slides and refine the lecture scripts in accordance with the ICAP framework, embedding guiding questions and constructive learning elements to foster cognitive engagement and higher-order thinking, while maintaining content consistency. The optimized instructional script is processed through a Speech Synthesizer, specifically, Google Text-to-Speech, to generate the lecture narration. The synthesizer modulates speech rate, intonation, and volume to ensure clarity and listening comfort. The resulting audio is then validated against the optimized script to guarantee precise alignment between text and narration, thereby eliminating discrepancies between spoken content and visual materials. In the final stage, the Video

Assembler synchronizes slide frames with corresponding audio segments and compiles them into a complete lecture video using the FFmpeg framework, producing a cohesive and pedagogically aligned output.

Experiment

Evaluation Setting

We employ GPT-4o-mini as the base model for the ICAP Evaluator. The system was evaluated using four primary metrics: (1) ICAP Assurance checks whether each video includes constructive questions or activities that align with the constructive level of the ICAP framework, implemented as interaction-eliciting instructional cues intended to encourage learners’ active reasoning, reflection, or self-explanation. (2) Generation Time Ratio is the total video generation time divided by the duration of the generated video. A lower value indicates faster generation. (3) Speech Consistency evaluates the match between generated speech and input script based on ROUGE-1 score (Lin, 2004). A higher value indicates a better consistency. (4) Slides Coverage is the ratio of input slides that appear in the final output video. The evaluation was conducted on five university-level courses comprising hundreds of slides, with Google NotebookLM used as the baseline.

Main Results

As shown in Table 1, the proposed system outperforms Google NotebookLM across all four metrics. Our system has a lower Generation Time Ratio, addressing the time-consuming nature of lecture production. Our system also has perfect Speech Consistency and Slides Coverage scores, indicating precise alignment with instructor scripts and complete use of provided materials, overcoming the issue of inconsistent or inaccurate content. Finally, ICAP Assurance is guaranteed in our system, ensuring the intentional integration of constructive learning prompts and addressing the lack of learning-science principles in existing tools.

Table 1.
Results of our system compared with NotebookLM. For Generation Time Ratio, the lower the better; for Speech Consistency and Slides Coverage Rate, the higher the better.

Method	Generation Time Ratio (↓)	Speech Consistency (ROUGE-1, ↑)	Slides Coverage Rate (% ↑)	ICAP Assurance*
Ours	0.29	100.00	100.00	Guaranteed
NotebookLM	0.91	19.82	39.81	Incidental Only

*ICAP Assurance refers to the intentional and guaranteed presence of constructive prompts. NotebookLM may produce constructive-style guiding questions, but these occur incidentally and are not controlled by an instructional design framework.

Discussion

The results show that controllability and instructional alignment improve the practicality of AI-generated lecture videos. The system consistently satisfied ICAP Assurance, produced accurate narration, preserved full slide coverage, and generated complete videos within a predictable time. These features address common issues in existing AI video tools and allow instructors to create pedagogically aligned materials with reduced effort. This supports more scalable MOOC development.

Future work will integrate a Learning Collaborator Agent to introduce interactive learning opportunities within the ICAP framework.

References

- Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., and Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review*, 37(2), 45.
- Chi, M. T., and Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219-243.
- Google, 2025. NotebookLM [Computer software]. <https://notebooklm.google/>
- Guo, P. J., Kim, J., and Rubin, R. (2014, March). How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the First ACM Conference on Learning at Scale* (pp. 41-50).
- Hollands, F. M., and Tirthali, D. (2014). MOOCs: Expectations and Reality. Full Report. Online Submission.
- Lin, C.-Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. *Workshop on Text Summarization Branches Out*, Post-Conference Workshop of ACL, Barcelona, Spain.
- Yao, H., Xu, W., Turnau, J., Kellam, N., and Wei, H. (2025). Instructional agents: LLM agents on automated course material generation for teaching faculties. *arXiv preprint arXiv:2508.19611*.



From Measurement to Action: A Learning Engineering Approach to AI-Powered Assessment for Human Power Skills Development

Yigal Rosen & Ilia Rushkin

AI-Enabled Assessment

Power Skills

Talent Development

As AI systems transform work at scale, human “power skills”—including creative thinking, communication, collaboration, leadership, analytical reasoning, productivity, and AI fluency—are increasingly critical to employability and long-term success. Yet these skills remain difficult to measure rigorously and to develop systematically. This paper is explicitly grounded in Learning Engineering and demonstrates how a nested Learning Engineering cycle—spanning challenge definition, creation, implementation, and investigation—can be embedded within a novel AI-enabled assessment system. Building on prior large-scale assessments of collaborative problem solving and creative thinking, adaptive learning research, and agentic AI, the Ignis AI PowerSkillsAssessment™ integrates evidence-centered design, generative AI, Bayesian proficiency estimation, and fairness-aware modeling into a unified operational framework. We describe the construct model, task design, scoring architecture, and early pilot results with professionals and

managers, illustrating how learning engineering can translate insights from cognitive science and psychometrics into scalable, actionable tools for talent development in the AI era.

Introduction

The accelerating deployment of AI systems in workplaces is reshaping the division of labor between humans and machines. Routine and narrowly defined tasks are increasingly automated, while enduring value is shifting toward human capabilities that complement rather than compete with AI: creative thinking, communication, collaboration, leadership, analytical reasoning, productivity, and the ability to work effectively with AI tools. International frameworks, such as the OECD's recent Skills Outlook 2025 and the World Economic Forum's Future of Jobs Report 2025, converge on the view that human power skills—such as creative thinking, adaptability, communication, and leadership—are central to future employability, lifelong learning, and innovation across sectors (OECD, 2025; World Economic Forum, 2025). Despite this recognition, most existing tools for assessing these capabilities rely on self-report questionnaires or small-scale divergent-thinking tests, which suffer from limited construct validity, social desirability bias, and poor alignment with real-world performance. There is a pressing need for assessment systems that can (a) capture how people actually perform in realistic, cognitively demanding situations, (b) provide interpretable, actionable evidence to support learning and development, and (c) scale across contexts, roles, and industries. The Ignis AI PowerSkillsAssessment™ was designed to address this need by integrating psychometrically grounded constructs, AI-mediated task environments, automated scoring, and Bayesian proficiency estimation into a unified system. This paper describes how learning-engineering principles guided the translation of research on power skills into an operational assessment platform that can support scalable and equitable talent development.

This work is situated within the field of Learning Engineering, which integrates learning sciences, human-centered design, data, and engineering practices to translate research insights into scalable, evidence-based learning systems. As articulated by Baker et al. (2022) and Goodell et al. (2023), Learning Engineering emphasizes iterative design, implementation, and evaluation grounded in real-world constraints. The Ignis AI PowerSkillsAssessment™ represents a concrete instantiation of this approach, operationalizing a nested Learning Engineering cycle within a broader AI-enabled platform to support the measurement and development of complex human skills (Thai et al., 2023).

Theoretical and Empirical Foundations

The design of the PowerSkillsAssessment™ draws on three interlocking strands of prior work. First, research on large-scale assessment of “hard-to-measure” social–cognitive skills, including the PISA 2015 Collaborative Problem Solving assessment, demonstrated that computer-driven conversational agents can emulate key aspects of human collaboration—such as establishing shared understanding, managing conflict, and coordinating action—within standardized digital environments (OECD, 2017; Rosen, 2014, 2015). Results showed that performance in human-to-agent settings correlated strongly with human-to-human collaboration while providing richer and more consistent behavioral data. Second, the PISA 2022 Creative Thinking assessment offered the first internationally validated framework and computer-based assessment for creative thinking in 15-year-olds, conceptualizing creative thinking as the ability to generate, evaluate, and improve ideas across multiple domains (OECD, 2024; Rosen, Stoeffler, & Simmering, 2020). This work moved beyond small-scale divergent-thinking tasks and self-report measures by using open-ended, scenario-based activities evaluated through rigorous scoring rubrics and psychometric modeling. Third, research on adaptive learning and proficiency estimation at Harvard and in large-scale online courses provided a foundation for Bayesian modeling of skill mastery and dynamic, evidence-based feedback. Projects such

as the Adaptive Learning Open Source Initiative (ALOSI) applied Bayesian knowledge-tracing and open adaptive engines in MOOCs to estimate learners' evolving mastery states across multiple skills and adjust instruction accordingly (Rosen et al., 2017, 2018). These strands—agentic AI in assessment, creative thinking measurement at scale, and Bayesian adaptive modeling—collectively informed the design of a power-skills assessment architecture that is both scientifically robust and operationally scalable.

Power Skills Framework and Construct Model

The system begins with a construct architecture that defines each skill through observable, evidence-aligned components. A construct model displaying major power skills—Communication, Collaboration, Creative Thinking, Leadership, Analytical Reasoning, Productivity, and AI Fluency. Each domain is further decomposed into subskills grounded in prior research and psychometric theory. For example, creative thinking is operationalized in terms of originality, flexibility, elaboration, and idea improvement, consistent with the PISA 2022 framework and creativity literature (Rosen et al., 2020; Kaufman & Beghetto, 2009). Communication includes clarity, audience awareness, structure, and tone; collaboration includes coordination, perspective-taking, and constructive feedback; leadership focuses on direction-setting, alignment, and support; analytical reasoning incorporates interpretation, inference, and argumentation; productivity includes goal-setting, prioritization, and follow-through; and AI fluency focuses on problem formulation, prompt strategy, and critical evaluation of AI-generated outputs. This construct model guides all downstream design decisions: item specifications, task environments, rating rubrics, scoring models, and reporting structures. It also provides a basis for evaluating construct validity by examining whether observed performance patterns align with theoretically expected relationships among subskills and domains (Messick, 1989; Mislevy, 2013; Totino & Kessler, 2024).

Task Design and AI-Assisted Content Generation

The assessment uses open-ended and scenario-based tasks that emulate realistic cognitive and interpersonal demands. Examples include: (a) Composing a response to a complex stakeholder email (Communication + Leadership); (b) Providing structured peer feedback on a teammate's proposal (Collaboration + Communication); (c) Generating multiple solutions to an ambiguous workplace challenge and refining one into a detailed plan (Creative Thinking + Analytical Reasoning); and (d) Evaluating AI-generated outputs for a given prompt, identifying limitations, and revising the prompt to improve the result (AI Fluency + Productivity). Task development follows evidence-centered design principles (Mislevy et al., 2003): for each subskill, designers specify the claim (what is being inferred), the evidence (what behaviors indicate proficiency), and the task features (what situations are likely to elicit that evidence). This structure ensures alignment between intended constructs and observable behaviors. To accelerate content development while preserving validity, the system uses a controlled generative-AI pipeline. Large language models are prompted with structured templates and constraints derived from the construct model and item specifications. Candidate tasks are generated, automatically screened for basic quality and redundancy, and then reviewed by human experts for construct alignment, clarity, and fairness. This approach allows coverage of multiple industries and roles while keeping humans in the loop for high-value decisions, such as ensuring that tasks truly elicit targeted subskills and are culturally and contextually appropriate.

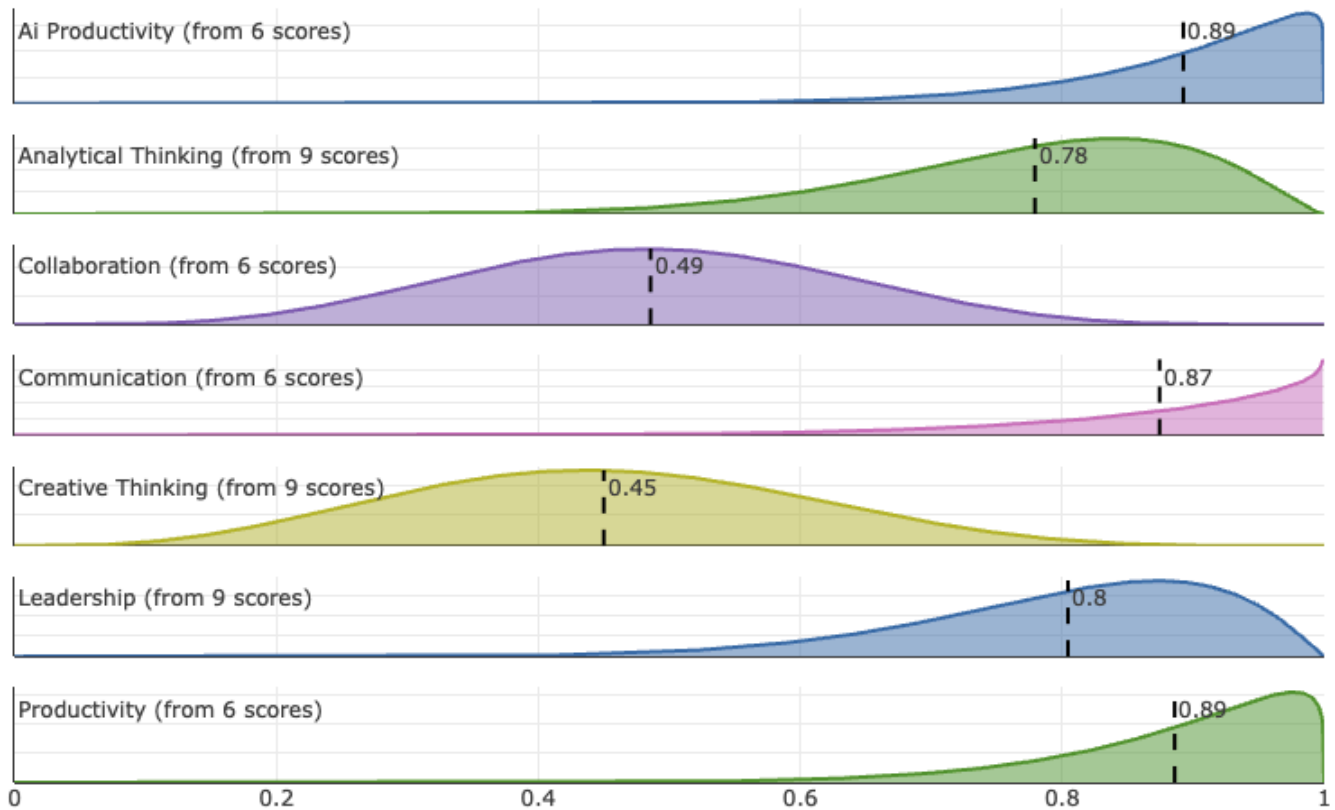
Scoring Architecture and Bayesian Proficiency Estimation

Responses are scored using a hybrid human-machine approach. For each task, a rubric defines performance levels for relevant subskills, with behavioral descriptors anchored in prior research (e.g., on creative thinking levels, collaborative problem-solving processes, and written communication quality). A portion of responses is double-scored by trained human raters to establish reliability and calibrate automated scorers. Automated scoring models leverage large language models fine-tuned or prompted with rubric exemplars. To mitigate construct-irrelevant variance and bias, prompt templates include rubric descriptors, and scoring outputs are audited for consistency and fairness across subgroups. Skill proficiency is estimated within a Bayesian latent-trait framework (e.g., multidimensional IRT or related models; von Davier, 2010). For each participant and each skill, the model maintains a posterior probability distribution over proficiency on a 0–1 scale. As responses

accumulate across tasks, the posterior updates, narrowing as evidence increases. Figure 1 illustrates an example posterior distribution.

Figure. 1.

Illustrative example for power skills estimation for an individual participant



The plots show the posterior probability distributions of the participant's proficiency level on each skill, measured on the scale from 0 to 1. The dashed vertical lines with numeric values indicate the means of those distributions, which serve as the reportable single-value estimates of proficiency. This probabilistic model aligns with learning-engineering principles by supporting both assessments (what is the current skill level?) and learning optimization (what additional evidence is most useful?).

Preliminary Empirical Evidence

Pilot studies with professionals and managers were conducted to evaluate the psychometric properties of the Ignis AI PowerSkillsAssessment™ and to illustrate how Learning Engineering connects design decisions to data-driven improvement across iterative cycles. Consistent with a nested Learning Engineering methodology, these pilots were used not only to validate measurement quality, but also to inform refinement of task design, scoring models, and proficiency estimation (Study I and Study II). Together, the studies focused on instrument reliability, factor structure, item performance, uncertainty modeling, and fairness across diverse populations.

Study I was conducted in a low-stakes setting with a diverse U.S. adult sample recruited via the Prolific platform (n = 353). Participants completed scenario-based tasks spanning creative thinking, communication, collaboration, leadership, analytical

reasoning, productivity, and AI fluency. Psychometric analyses indicated strong internal consistency across domains, with Cronbach's α ranging from .82 to .91—well within accepted ranges for short, performance-based assessments involving open-ended and ordered multiple-choice responses. Exploratory and confirmatory factor analyses supported the hypothesized multidimensional structure of the Power Skills framework, with each domain loading strongly on its intended latent factor and moderate inter-factor correlations, consistent with related but distinct constructs. Item-level analyses demonstrated a healthy spread of difficulty and discrimination parameters, and standard workforce item-selection criteria resulted in a 96% acceptance rate in this preliminary pilot. Qualitative inspection of constructed responses further revealed meaningful differences between higher- and lower-performing participants, particularly for indicators of originality, adaptability, and integrative decision-making in creative thinking and leadership tasks.

Study II extended this work in a higher-stakes, applied context with employees from two technology organizations ($n = 36$), representing roles in product management, software development, UX/UI design, and senior leadership. This pilot included an expanded task set and was designed to evaluate psychometric calibration, participant experience, and the behavior of the Bayesian proficiency estimation model under real-world conditions. AI-assisted scoring models, supported by human-in-the-loop validation, were applied to constructed and ordered responses, and posterior proficiency distributions were generated for each skill. Results indicated robust reliability across domains (average $\alpha = .85$), and confirmatory factor analyses again supported the framework's skills and subskills structure. Item information curves showed optimal precision in mid- to high-proficiency ranges, which is particularly important for identifying emerging leaders and supporting targeted development. Approximately 75% of items met workforce-grade psychometric criteria in this applied setting, providing clear guidance for ongoing optimization of the assessment.

Across both studies, Bayesian proficiency estimation enabled explicit modeling of uncertainty, with posterior distributions narrowing as additional evidence was incorporated through more items, multi-tagged tasks, and supplemental data sources. This approach allowed proficiency scores to be reported alongside credible intervals, reinforcing interpretability and appropriate use of results. Convergent validity analyses comparing performance-based estimates with self-report and peer-review indicators showed moderate correlations, consistent with prior research suggesting that behavioral assessments capture related but non-redundant variance compared to perception-based measures. Collectively, these findings provide promising preliminary evidence that the Ignis AI PowerSkillsAssessment™ can generate reliable, fair, and developmentally informative measures of human power skills in both low- and higher-stakes contexts, while exemplifying how Learning Engineering principles support continuous improvement from empirical evidence to system refinement.

Learning Engineering in Action: From Insights to Implementation

The design and refinement of the PowerSkillsAssessment™ follow a learning engineering cycle: (a) Insight generation from prior research and large-scale implementations (e.g., PISA 2015/2022, adaptive learning in MOOCs) informs the construct model and task design; (b) Prototyping of tasks, scoring prompts, and proficiency models is carried out in small-scale pilots; (c) Data collection and analysis through pilots provide empirical evidence on reliability, validity, item performance, fairness, and user experience; (d) System redesign uses this evidence to revise tasks, rubrics, scoring procedures, and proficiency estimation, including the refinement of AI prompts and filters; and (e) Implementation in real-world contexts (e.g., workforce development programs, higher education settings) produces additional data to inform both product evolution and research questions.

This iterative process exemplifies learning engineering: it integrates educational science, design, AI, and data systems to produce tools that are not only technically sophisticated but also usable, interpretable, and aligned with real-world constraints. Insights from pilots directly affect implementation decisions, such as how many tasks per skill are required to achieve acceptable uncertainty thresholds, which task types best differentiate proficiency levels, and how to balance efficiency with depth of evidence.

Discussion and Future Directions

The early evidence suggests that an AI-powered, performance-based assessment of human power skills is technically feasible, psychometrically defensible, and practically applicable in talent development contexts. The combination of construct modeling, AI-assisted task generation, rubric-grounded scoring, Bayesian proficiency estimation, and fairness auditing constitutes a coherent learning-engineering pipeline.

Several directions for future work are particularly promising. First, longitudinal studies are needed to evaluate how proficiency estimates change over time in response to targeted learning experiences, including AI-supported coaching and practice. Second, predictive validity should be examined by linking power skills profiles to real-world outcomes such as talent development, job performance, promotion, and engagement. Third, the interaction between AI fluency and other power skills merits deeper investigation: for example, how individuals who effectively leverage AI tools differ in their creative output and collaboration quality compared to those who do not.

From a learning engineering perspective, continued work is required to refine the feedback layer—how to transform posterior distributions and profile patterns into actionable, personalized recommendations for learners, coaches, and organizations. This includes experimentation with different feedback formats, visualizations, and integration points within learning and talent platforms.

References

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*, 3(1). <https://doi.org/10.1037/tmb0000058>
- Craig, S. D., Avancha, K., Malhotra, P., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a nested learning engineering methodology to develop a team dynamic measurement framework for a virtual training environment. In *ICICLE 2024 Conference Proceedings: Solving for Complexity at Scale* (pp. 115–132). <https://doi.org/10.59668/2109.21735>
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. *Journal of Military Learning, Conference Edition*. <https://www.armyupress.army.mil/Journals/Journal-of-Military-Learning/Journal-of-Military-Learning-Archives/Conference-Edition-2023-Journal-of-Military-Learning/Engineering-at-a-Glance/>
- Goodell, J., & Kolodner, J. (Eds.). (2023). *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond*. Routledge.
- Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four C model of creativity. *Review of General Psychology*, 13(1), 1–12.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
- Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine*, 178(10), 107–114.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.
- OECD. (2017). *PISA 2015 results (Volume V): Collaborative problem solving*. OECD Publishing. <https://www.oecd.org/content/dam/oecd/en/publications/reports/2017/11/pisa-2015-results-volume->

v_g1g83e07/9789264285521-en.pdf

- OECD. (2024). PISA 2022 results (Volume IV): Creative thinking. OECD Publishing.
https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/pisa-2022-results-volume-iv_125a58b3/5a849c2a-en.pdf
- OECD. (2025). OECD Skills Outlook 2025: Building the Skills of the 21st Century for All. OECD Publishing.
https://www.oecd.org/en/publications/oecd-skills-outlook-2025_26163cd3-en.html?utm_source=chatgpt.com
- Page, E. B. (2000). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 68(2), 127–142.
- Rosen, Y. (2014). Comparability of conflict opportunities in human-to-human and human-to-agent online collaborative problem solving. *Technology, Knowledge and Learning*, 19(1–2), 147–174.
- Rosen, Y. (2015). Assessing students in human-to-agent settings to inform collaborative problem-solving learning. *Journal of Educational Measurement*, 52(4), 345–370.
- Rosen, Y., Stoeffler, K., & Simmering, V. (2020). Imagine: Design for creative thinking, learning, and assessment in schools. *Journal of Intelligence*, 8(2).
- Rosen, Y., Jaeger, G., Newstadt, M., Bakken, S., Rushkin, I., Dawood, M., & Purifoy, C. (2023). A multidimensional approach for enhancing and measuring creative thinking and cognitive skills. *International Journal of Information and Learning Technology*, 40(4), 334–352.
- Rosen, Y., Rushkin, I., Ang, A., Fredericks, C., Tingley, D., & Blink, M.-J. (2017). Designing adaptive assessments in MOOCs. In *Proceedings of the Fourth ACM Conference on Learning at Scale*.
- Rosen, Y., Rushkin, I., Ang, A., Munson, L., Lopez, G., Tingley, D., & Weber, G. (2018). The effects of adaptive learning in a massive open online course on learners' skill development. In *Proceedings of the Fifth ACM Conference on Learning at Scale*. <https://dl.acm.org/doi/10.1145/3231644.3231651>
- Thai, K. P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. (2023). Learning engineering is human-centered. In J. Goodell (Ed.), *The learning engineering toolkit* (pp. 83–124). Routledge.
- Totino, L., & Kessler, A. (2024). "Why did we do that?" A systematic approach to tracking decisions in the design and iteration of learning experiences. *Journal of Applied Instructional Design*, 13(2).
- von Davier, M. (2010). *Statistical models for test equating, scaling, and linking*. Springer.
- World Economic Forum. (2025). The Future of Jobs Report 2025. World Economic Forum.
https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf?utm_source=chatgpt.com



Implementing Concept Instruction via MCP Server

Thor A. Anderson

AI-powered learning

Concept teaching

critical attributes

Educational Technology

Instructional Design

Large Language Models

learning engineering

Model Context Protocol (MCP)

This applied learning engineering research demonstrates the operationalization of Merrill & Tennyson's (1977) concept teaching framework through AI-powered tools. An MCP (Model Context Protocol) server was developed implementing five sequential tools that decompose concept lesson creation into theory-aligned stages: concept definition, attribute analysis, example generation, practice activity creation, and lesson publication. The system encodes M&T's research-based prescriptions as structured prompts, validation algorithms, and schema-driven workflows. Key findings reveal structural constraints with input/output schemas are essential for maintaining theoretical fidelity—unconstrained LLM generation produced inconsistent, theoretically misaligned outputs despite natural language descriptions of instructional principles. The architecture combining LLM semantic capabilities with seven algorithmic validators suggests a promising pattern for educational AI: leverage LLMs for content generation

*while enforcing learning science principles through external validation.
Future work will embed these concept lessons within problem-based
learning environments for just-in-time instruction.*

Introduction

The proliferation of AI code generation tools has fundamentally disrupted traditional assessments in programming education —when students can prompt ChatGPT or GitHub Copilot to produce syntactically correct code, the ability to write code no longer serves as a reliable proxy for conceptual understanding. This shift elevates the importance of concept learning: educators must now focus on whether learners truly comprehend “why” code works, not merely whether they can produce it. Merrill and Tennyson’s seminal work “Teaching Concepts: An Instructional Design Guide” (1977) synthesized six empirical studies into prescriptive principles for concept instruction. Despite years of research establishing clear instructional prescriptions, these principles have remained challenging to systematically implement at scale in contemporary digital learning environments.

The emergence of large language models (LLMs) and the Model Context Protocol (MCP) presents both opportunities and challenges for learning engineering. While LLMs demonstrate impressive capacity for semantic analysis and code generation, their application to instructional design often produces inconsistent results when left unconstrained. As Merrill et al. (1996) emphasize, merely generating content about concepts does not constitute pedagogically sound instruction. Efforts to use AI for generating “online textbooks” frequently conflate information delivery with effective concept teaching, ignoring research-based prescriptions about example selection, attribute identification, and boundary case representation.

Recent developments in structured AI tool orchestration through MCP suggest a novel approach: encoding Merrill & Tennyson’s research-based prescriptions as formal validation algorithms, structured prompts, and composable tools that leverage LLM capabilities while maintaining theoretical fidelity. This approach imposes external constraints on LLM outputs through input/output schemas and automated quality checking against M&T’s empirically derived instructional principles.

Methods

An MCP server was developed as a web application, implementing five sequential tools that decompose concept lesson creation into discrete, theory-aligned stages:

Define concept	Analyze concept	Generate examples	Create practice activities	Publish lesson

The system operates in two modes to support different deployment contexts: Prompt Mode returns structured prompts for execution by Claude Desktop’s selected model. LLM Mode makes direct API calls to Ollama (local model) or OpenAI (remote) for autonomous operation.

This hybrid architecture was essential for practical deployment, as early attempts to allow unconstrained LLM generation produced highly inconsistent outputs. The structured input/output schemas enforced by MCP tools proved critical for

maintaining theoretical fidelity. Merrill & Tennyson's research-derived prescriptions were operationalized through three complementary mechanisms:

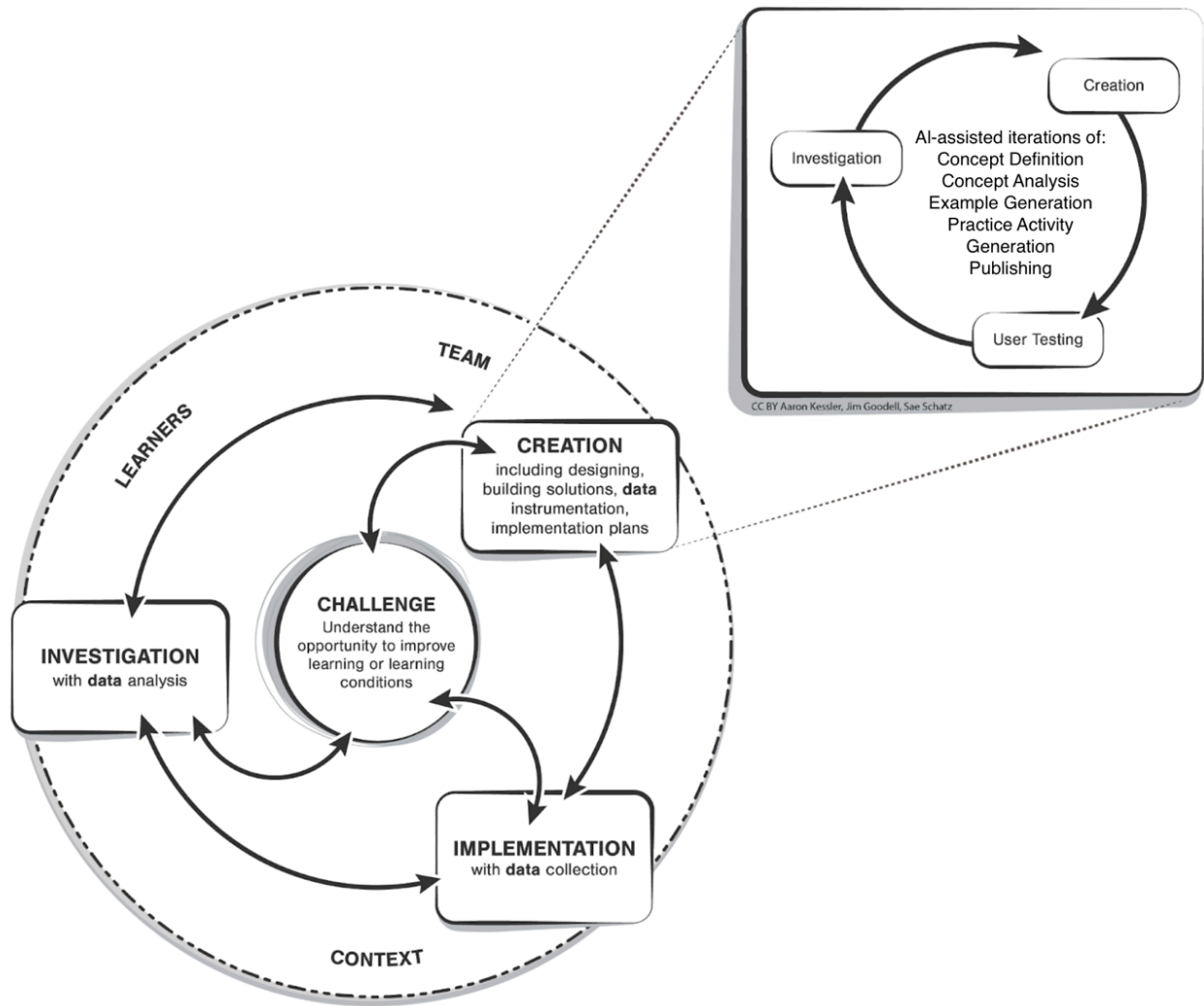
1. **Structured System Prompts.** Each tool receives domain-specific prompts encoding M&T principles as explicit constraints.
2. **Validation Algorithms.** Algorithmic validators ensure outputs satisfy theoretical requirements: they verify that all positive examples contain ALL critical attributes and that negative examples are missing at least one. Diversity scores (0-1) are computed measuring variation on non-critical features to prevent stereotype bias. Balanced example ratios (approximately 50-70% positive examples to negative examples) are computed. "Near-miss" negative examples (missing exactly one critical attribute) are validated. Theory-based teaching warnings (overgeneralization risks, misconceptions, boundary ambiguities) are generated. Quality scores with actionable improvement recommendations are provided and insufficient variable attribute variations are flagged.
3. **Schema-Driven Outputs.** Zod (a TypeScript validation library) schemas enforce structured data flow between tools to prevent LLMs from inventing their own interpretation of what constitutes "good" concept instruction.

This implementation follows a nested Learning Engineering approach (Craig et al., 2025), where the concept lesson creation and validation tools iterate on the production of concept lesson content inside the broader instructional design cycle as shown in Figure 1. It applies the IEEE ICICLE definition of Learning Engineering by coupling human-centered design with data-informed, iterative improvement (Goodell et al., 2022).

Figure 1.

Learning Engineering Process With Adapted Depiction of Nested Creation Phase.

Note. Adapted from Craig et al. (2025). CC BY.



The AI-assisted algorithmic content creation and validation automate the nested cycle portion of the Learning Engineering process. This greatly accelerates the creation-phase iteration by shifting portions of the iteration that typically require expert human judgement from creation to review and testing. Each MCP tool run challenges the LLM's outputs against Merrill & Tennyson constraints, applies structured fixes via schemas, and investigates compliance before passing results downstream. This operationalizes Baker et al.'s (2022) call for computational methods that enable faster experimentation and iteration, turning learning-science prescriptions into standardized, reusable components. This greatly accelerated experimentation and iteration allowed the author to produce a set of three online courses for the current semester. Each course is structured to have seven "sprint" sessions in the semester with each sprint having between 33 to 36 topics. Each sprint tutorial is linked to concept lessons which required the generation of 281 concept lessons (each lesson containing approximately 8,000 lines of HTML) within an approximately 4-week time span; a task that is clearly only possible with the help of generative AI and agentic tooling.

Results

The system successfully operationalized Merrill & Tennyson's key instructional prescriptions into executable code and structured prompts that achieve critical attribute coverage, example diversity, example balance, boundary representation, and tool chain coherence.

Given the system's current development stage as a practitioner-focused tool, evaluation focused on:

1. Architectural validation: Verification that M&T principles are correctly encoded in prompts, schemas, and validation algorithms.
2. Tool chain coherence: Testing that outputs from one tool properly feed into subsequent tools (e.g., `define concept` → `analyze concept` → `generate examples`).
3. Practical usability: Observation of lesson generation workflows in Claude Desktop, including error handling and iterative refinement.
4. Code quality review: Manual inspection of generated code examples for production quality, realistic contexts, and pedagogical clarity.

The focus on architectural fidelity rather than large-scale empirical testing reflects the system's status as an applied learning engineering prototype designed to demonstrate feasibility rather than claim efficacy at scale.

Discussion

This work demonstrates that classical instructional design theory can be systematically encoded into AI tool chains that maintain theoretical fidelity while leveraging LLM semantic capabilities. The success of the hybrid architecture—LLM semantic analysis combined with algorithmic validation—suggests that M&T's principles possess sufficient precision to serve as formal specifications for automated instructional systems.

The most significant finding was that structural constraints are essential for theoretical fidelity. Allowing LLMs to generate lessons from unstructured prompts produced inconsistent and theoretically misaligned outputs. Enforcing strict input/output schemas through MCP tool interfaces and careful wording of system prompts achieved reliable adherence to instructional design principles.

Following Merrill's (2002) "First Principles of Instruction", future work will investigate embedding concept lessons within larger problem-based scenarios. This integration addresses Merrill et al.'s (1996) core argument that effective learning occurs when information supports problem-solving rather than existing as isolated content.

References

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*, 3(1). <https://doi.org/10.1037/tmb0000058>
- Craig, S. D., Avancha, K., Malhotra, P., Gorman, J. C., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a nested learning engineering methodology to develop a team dynamic measurement framework for a virtual training environment. *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale*. <https://doi.org/10.59668/2109.21735>

- Goodell, J., Kolodner, J., & Kessler, A. (2022). Learning engineering applies the learning sciences. In J. Goodell & J. Kolodner (Eds.), *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond* (pp. 47-82). Routledge.
- Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development*, 50(3), 43-59.
- Merrill, M. D., Drake, L., Lacy, M. J., Pratt, J., & the ID2 Research Group. (1996). Reclaiming instructional design. *Educational Technology*, 36(5), 5-7.
- Merrill, M. D., & Tennyson, R. D. (1977). *Teaching concepts: An instructional design guide*. Educational Technology Publications.



L2-French Learners and Generative AI (GenAI): Challenges, Needs, and Design Guidelines

Jiachen Gong, Ishrat Ahmed, Fan Yu, Tracy Arner, & Danielle S. McNamara

In this study, we examined Second Language (L2)-French learners' experiences, challenges, perceptions, and expectations of GenAI tools for French learning by conducting a survey. Twenty-two L2-French learners from a large public university completed the survey. Results showed that students actively seek additional learning resources outside of class, yet continue to struggle with memorizing grammar and vocabulary. Although most participants reported general familiarity with GenAI tools, few had applied them to French learning. Students were interested in a customized GenAI tool that could provide personalized practice as a supplemental resource. These findings offer valuable guidance for the design and development of a GenAI-assisted French Learning system that delivers personalized grammar and vocabulary practices aligned with learners' needs.

Jiachen Gong¹, Ishrat Ahmed¹, Fan Yu¹, Tracy Arner¹ and Danielle S. McNamara¹

¹ Learning Engineering Institute, Arizona State University

Abstract.

Keywords: L2-French learning, GenAI in Education, Learner-Centered Design

Introduction

Technology is significantly transforming second language acquisition (SLA) by leveraging innovative tools that can enhance the educational experience (Chun et al., 2016). In this study, we adopt a learning engineering approach, situating our work within the Challenge phase of the learning engineering process, where the goal is to understand the learning problem in context and generate evidence to guide subsequent design decisions (Kessler et al., 2023). Innovative tools powered by Generative AI (GenAI) (e.g., ChatGPT, Gemini, Llama, etc.) have opened new avenues for personalized and interactive learning environments (Baskara & Mukarto, 2023). In the context of SLA, these developments build on a history of leveraging natural language processing (NLP) in SLA personalization (examples include Amaral et al., 2011; Heift et al., 2010). Therefore, adapting from traditional NLP methods to the GenAI models is natural and well-motivated. Research shows that these AI-driven technologies have been used in SLA to provide real-time feedback, personalized tutoring, and interactive conversational practice, displaying the potential to enhance SLA experience (Barrot, 2024; Bin-Hady et al., 2025; Sa'ad et al., 2023; Woo & Choi, 2021). Despite the growing body of research on AI in SLA, a notable gap remains in understanding the learner- and context-specific evidence needed to justify design decisions for AI-supported French learning systems. Existing SLA research overwhelmingly centers on English (Law, 2024; Morrison, 1996). This imbalance is further reinforced by the fact that most large language models (LLMs) are predominantly trained on English-dominant corpora (Tao et al., 2024), raising questions about their validity and reliability when applied to SLA contexts beyond English. From a learning engineering perspective, this gap underscores the need for early-stage, context-sensitive needs analyses to guide design decisions (Goodell et al., 2023; Thai et al., 2023).

Accordingly, this study represents an initial, human-centered phase within a broader learning engineering cycle focused on developing a GenAI-assisted French learning system. To support L2-French learners, we first need to understand how students learn French outside the classroom, their experiences with GenAI tools, the tools they currently use, the challenges they encounter, their perspectives on GenAI, and their preferences for using GenAI in language learning. Gaining knowledge on these perspectives, the research can assist in developing AI-driven solutions that align with students' needs and may enhance the effectiveness of French language education. To address this, in this paper, we surveyed 22 participants to understand user needs, challenges, and expectations related to using GenAI tools for French learning with the goal of informing subsequent system design and iteration.

Method

Measurement

We developed a questionnaire to collect information about students' backgrounds, French learning experiences, and perceptions of GenAI tools to inform the challenge of designing a GenAI-assisted French learning system. The instrument consisted of five sections: 1) Student demographics included academic standing, enrollment status (full-time/part-time), native language, and major; 2) French learning experiences included asking students which French class they enrolled in, what strategies they used to learn French outside of class, and what challenges they encountered while learning French; 3) Usage of GenAI tools included whether they had used GenAI and, if so, how they applied these tools to French learning; 4) Perception of GenAI tools, included asking students' views on the effectiveness or limitations of GenAI tools for supporting French learning; and finally, 5) Needs and expectations of GenAI tools included asking students which areas of French learning they would like

GenAI tools to support and what features they would find beneficial. The questionnaire included multiple-choice items, multiple-select checkbox questions, and open-ended questions to elicit both quantitative and qualitative data.

Participants

A total of 29 French students at a large, public university provided consent to participate in the survey. Seven respondents were excluded from analysis as they completed only the demographic section, resulting in a final sample of 22 participants. Most participants ($n = 20$) were enrolled in a lower-division French course. The sample included six freshmen, five sophomores, three juniors, six seniors, and two students who selected “other”. The majority of participants were full-time students ($n = 19$). English was the predominant native language among participants ($n = 19$). Participants represented a diverse range of academic majors: six in the Humanities (e.g., French, History), four in the Natural Sciences (e.g., Biochemistry, Chemistry), six in the Social Sciences (e.g., Anthropology, Psychology), two in the Arts, and four in other disciplines.

Procedure and Analysis

The survey was distributed twice, once in Fall 2024 and again in Spring 2025, to students enrolled in French 101, 102, 201, and 202. This study received approval from the Institutional Review Board (IRB), and all participants provided informed consent before completing the survey. Participation was voluntary; students could decline to answer any questions.

Frequency counts were calculated to summarize participants’ responses to multiple-choice and multi-select questions. Thematic coding was performed on open-ended responses by two researchers and agreed upon to identify recurring learner needs, challenges, and contextual constraints relevant to subsequent design decisions (Vaughn & Turner, 2016).

Results

To better understand the learning challenge faced by L2-French learners in this context, the following results summarize evidence collected to inform design considerations for a GenAI-assisted French learning system. Findings are organized under four categories that reflect key inputs to the learning engineering process for system design: 1) French learning experiences, 2) Use of GenAI tools, 3) Perceptions of GenAI tools, and 4) User needs and expectations for a GenAI tool designed to support French learning.

French learning experiences

Students reported a variety of methods used outside of class to support their French learning. Among the 15 respondents, most engaged with native French content (e.g., books, movies, music) ($n = 11$) and language-learning apps such as Duolingo ($n = 8$). Other approaches included speaking with native French speakers ($n = 3$), watching videos or podcasts ($n = 2$), using flashcards ($n = 2$), utilizing supplemental French textbooks or dictionaries ($n = 2$), and private tutoring ($n = 1$). Students also reported several challenges in learning French. Among the 19 respondents, 12 mentioned memorization as a major challenge, with specific struggles related to remembering grammar rules ($n = 4$), verb conjugation ($n = 4$). Additional challenges included understanding spoken French ($n = 4$), limited speaking practices ($n = 3$), pronunciation ($n = 3$), lack of motivation ($n = 2$), overwhelming class content ($n = 2$), insufficient learning resources ($n = 1$), and inadequate guidance in planning their learning ($n = 1$).

Use of GenAI Tools

We asked participants whether they had used GenAI tools such as ChatGPT and Gemini. Although most students ($n = 17$) had used GenAI, more than half ($n = 12$) reported that they had not used these tools specifically for learning French. Follow-up

questions were asked to five students who had used GenAI for French learning. Among these students, four reported using ChatGPT or its variants, and 1 reported using Gemini.

Students most commonly reported using GenAI to prepare for exams ($n = 4$) and to complete homework or assignments ($n = 3$). Regarding how they used GenAI to learn French, most described using it as a tutor ($n = 4$) or as a practice platform ($n = 3$). When asked which task(s) they used GenAI for, students indicated writing ($n = 4$), reading ($n = 4$), grammar support ($n = 4$), speaking ($n = 3$), vocabulary learning ($n = 3$), and listening ($n = 2$).

Perception of GenAI Tools

We asked students about their perceptions of the effectiveness of GenAI tools for French learning. When asked whether GenAI tools could be effective, eight students responded that they are unsure, six responded yes, two responded no, and the remaining students did not respond. Students who responded yes or unsure ($n = 14$) were then asked how GenAI could support their learning. The most reported benefit was generating practice exercises ($n = 12$), followed by providing explanations to their questions ($n = 11$), giving feedback ($n = 8$), answering questions ($n = 6$), and supporting self-directed learning ($n = 6$). Students who responded no or not sure ($n = 10$) were also asked in what ways GenAI may be ineffective. The most common concerns were that GenAI may introduce academic dishonesty ($n = 8$) and provide inaccurate answers or confusing explanations ($n = 7$). Additional concerns included the limited usefulness of GenAI materials ($n = 5$), the ease of obtaining answers without effort ($n = 5$), and the time required to use GenAI ($n = 1$).

When asked whether they would be interested in using GenAI tools more frequently if they were tailored to French learners, 10 students responded yes, four were unsure, two responded no, and the others did not respond. We further asked students to explain their interest or lack of interest in customized GenAI French learning tools; 10 students responded with varied reasons. The primary reason for interest was the potential for personalized learning experiences ($n = 6$), including 1-on-1 tutoring. One student indicated they “would be more successful if I was able to learn 1-on-1 more often, rather than in a group,” in part because they would feel “significantly less nervous speaking with [GenAI].” Moreover, participants also indicated the need for more personalized and relevant practices. One student expressed the need to “include more examples of the topic we are working on”, and another indicated practices aligned to the concepts they care about and their learning experiences “would make it much more convenient.”

In contrast, the primary negative comments expressed by participants focused on biased GenAI outputs, ethical concerns, and environmental impacts. One student indicated that it was “too easy for generative AI to start spouting false statements,” and another indicated a concern about “AI replacing human input and generating incorrect and biased information.” Finally, another student noted they were “uncomfortable using AI because of its negative energy and environmental impacts”. Additionally, six students indicated some hesitation about AI tools. For example, one that viewed the GenAI tool as “a supplementary tool, not a main method of learning.” Another student said, “It will not be as effective as a human teacher; it is the next best option.”

Needs and Expectations of a French-Learning GenAI tool

We first asked students which areas of French learning they most needed support from a GenAI tool. Among the 19 respondents, 12 students identified grammar as their primary area of need. Students frequently described challenges related to verb conjugation; for example, one student noted difficulty with “distinguishing the uses of different tenses and remembering how to conjugate them.” Others emphasized struggles with complex grammatical structures, such as “making sure tenses and gender/plurality agreements are accurate.” In addition to grammar, five students reported needing support with pronunciation. One student expressed a desire for a tool that could “tell me how I’m forming my mouth to get the sounds right.” Additional areas of need included vocabulary ($n = 3$), speaking ($n = 2$), cultural context ($n = 2$), listening ($n = 1$), and learning planning ($n = 1$).

We further asked students what features they would include if they were to design a GenAI tool for French learning. Among the 18 responses, the most frequently mentioned feature was practice questions ($n = 7$). In particular, students wanted personalized practice opportunities aligned with their current levels of performance and class goals. For instance, one student wanted the tool to “generate exercises on things I mess up with most,” while another requested opportunities to “practice phrases and words that are used in the class.” Students also expressed interest in detailed explanations of grammar, vocabulary, or pronunciation ($n = 5$). One student described a desire for “grammar practice with detailed feedback and explanations,” while another indicated that the tool should “tell me how my sentences are wrong and how I should fix them.” Other desired features included integration with other media or software ($n = 5$), conversational practice with the GenAI ($n = 3$), a peer collaboration tool ($n = 1$), contextualized learning ($n = 1$), and listening practice ($n = 1$).

Discussion and Conclusion

In this paper, we examined learning challenges faced by L2-French learners in the context of emerging GenAI tools through a user needs analysis. Guided by a learning engineering approach (Kessler et al., 2023), we characterized learner needs, constraints, and contextual factors to inform the design of a GenAI-assisted French learning system.

Our findings primarily indicate that students actively seek supplemental materials beyond class, particularly to address difficulties with grammar and vocabulary memorization, which were areas identified as the greatest need for GenAI support. These results suggest prioritizing grammar- and vocabulary-focused assistance, potentially incorporating evidence-based strategies such as spaced repetition (e.g., Settles & Meeder, 2016). Second, students viewed GenAI’s primary value as generating personalized practice aligned with their learning progress and course content, with practice generation and adaptivity emerging as the most desired features. Last but not least, students raised concerns about academic dishonesty, inaccurate information, replacing human tutors, and environmental impact, indicating awareness of GenAI’s broader societal and ethical implications. This implies that the system must consider ethics: ensure accuracy, protect academic integrity, minimize environmental impact, and serve as a complement to—rather than a replacement for—human instruction.

In conclusion, the results of this study identified learner-centered challenges that may facilitate the design of AI-assisted French learning systems. While limited by sample size and potential interpretive bias, the findings provide rich insights into L2-French learners’ experiences and expectations in the era of GenAI. Future work will focus on designing, developing, and evaluating a multi-agent AI-assisted French learning system that delivers personalized grammar and vocabulary practice grounded in these findings.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

- Amaral, L., Meurers, D., & Ziai, R. (2011). Analyzing learner language: towards a flexible natural language processing architecture for intelligent language tutors. *Computer Assisted Language Learning*, 24(1), 1–16. <https://doi.org/10.1080/09588221.2010.520674>
- Barrot, J. S. (2023). ChatGPT as a Language Learning Tool: An Emerging Technology Report. *Technology, Knowledge, and Learning*, 29(29). <https://doi.org/10.1007/s10758-023-09711-4>

- Baskara, R., & Mukarto, M. (2023). Exploring the Implications of ChatGPT for Language Learning in Higher Education. *IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics)*, 7(2), 343–358. <https://doi.org/10.21093/ijeltal.v7i2.1387>
- Chun, D., Kern, R., & Smith, B. (2016). Technology in Language Use, Language Teaching, and Language Learning. *The Modern Language Journal*, 100(S1), 64–80. <https://doi.org/10.1111/modl.12302>
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. *Journal of Military Learning, Conference Edition*, 46–59.
- Heift, T. (2010). Developing an Intelligent Language Tutor. *CALICO Journal*, 27(3), 443–459. <https://doi.org/10.11139/cj.27.3.443-459>
- Kessler, A., Craig, S. D., Goodell, J., Kurzweil, D., & Greenwald, S. W. (2023). Learning engineering is a process. In J. Goodell & J. L. Kolodner (Eds.), *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond*. Routledge. <https://doi.org/10.4324/9781003276579>
- Morrison, L. (1996). Talking About Words: A Study of French as a Second Language Learners' Lexical Inferencing Procedures. *Canadian Modern Language Review*, 53(1), 41–75. <https://eric.ed.gov/?id=EJ535873>
- Sa'ad, A., Abdelhadi Alzyoud, AlShorman, O., & Basim Al-Absi. (2023). AI-driven Technology and Chatbots as Tools for Enhancing English Language Learning in the Context of Second Language Acquisition: A Review Study. *International Journal of Membrane Science and Technology*, 10(1), 1209–1223. <https://doi.org/10.15379/ijmst.v10i1.2829>
- Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1. <https://doi.org/10.18653/v1/p16-1174>
- Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9). <https://doi.org/10.1093/pnasnexus/pgae346>
- Thai, K. P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. L. (2023). Learning engineering is human-centered. In J. Goodell & J. L. Kolodner (Eds.), *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond*. Routledge. <https://doi.org/10.4324/9781003276579>
- Vaughn, P., & Turner, C. (2016). Decoding via coding: Analyzing Qualitative Text Data through Thematic Coding and Survey Methodologies. *Journal of Library Administration*, 56(1), 41–51. <https://doi.org/10.1080/01930826.2015.1105035>
- Wagdi Bin-Hady, Abdu Al-Kadi, Abduljalil Nasr Hazaea, & Kaid, J. (2023). Exploring the dimensions of ChatGPT in English language learning: a global perspective. *Library Hi Tech*, 43(4-5). <https://doi.org/10.1108/lht-05-2023-0200>
- Woo, J. H., & Choi, H. (2021). Systematic Review for AI-based Language Learning Tools. *ArXiv:2111.04455 [Cs]*. <https://arxiv.org/abs/2111.04455>



Learning-By-Explaining with Generative AI: A Pilot Implementation in Introductory Biology

Paul N. Ulrich, Kyra McNaughton, Ajibola Salami, & Kathryn S. McCarthy

Design Implementation Framework

Generative AI

Learning-by-explaining

STEM

Explanation activities promote learning but are difficult to scale. This project applied the Design Implementation Framework (DIF) in a nested Learning Engineering cycle to investigate how generative AI can support learning-by-explaining in introductory biology. In an experimental study, students were randomly assigned to view a brief video about explaining or an extended version that also modeled how to use ChatGPT to craft and improve explanations. The extended instruction resulted in longer exchanges and more causal language in the chats. Findings illustrate DIF's value for guiding AI integration and instruction that help students to leverage AI as a tool for active learning.

Introduction

Explanation is a powerful strategy for promoting conceptual understanding. However, students struggle to produce high-quality explanations without feedback. Instructors can scaffold this process, but the realities of high-enrollment courses make offering explanation activities and feedback challenging. Generative AI presents a promising path for scale but introduces its own set of barriers. On one hand, generative AI-based chatbots can foster active engagement and personalized support. On the other hand, casual use may yield cognitive disengagement in which students let the bot “do the thinking for them”. With this in mind, this work applies learning engineering to explore generative AI-supported explanation tasks in an introductory biology course.

Within this broader learning engineering effort, we employed the Design Implementation Framework (DIF; McCarthy et al., 2020) as a nested cycle (Craig et al., 2025). DIF includes five stages: defining and evaluating the problem, ideation, user experience/design, experimental evaluation, and implementation and feedback. Although similar to other frameworks, DIF emphasizes experimentation and co-design, which align with learning engineering’s focus on human-centered design (Thai et al., 2023). This project reflects a partnership across a learning sciences lab, biology instructor and the university Center for Excellence in Teaching, Learning and Online Education (CETLOE).

Method

Explanation activities were developed and implemented in two sections of introductory biology for non-majors. Courses met in-person, but the explanation activities were conducted asynchronously as homework.

Following DIF, we first defined the problem of using generative AI to support explanation strategies in large-enrollment courses and ideated solutions guided by the ICAP framework (Chi & Wylie, 2014) to promote constructive and interactive engagement. Although we considered creating a fully customized chatbot, our prior experiences suggested students tend to default to the general ChatGPT interface. Thus, we focused on designing instructional supports to help students use ChatGPT effectively and transfer these skills to new contexts. In the user experience/design phase, we developed a ChatGPT bot that collected basic student information and exported transcripts in .json format. The bot was piloted and refined with the CETLOE support team and research assistants.

We ran an implementation in Week 2 of the semester in which students completed a learning-by-explaining activity on transfer of energy in food chains (i.e., Trophic Pyramids and the Rule of 10), with minimal guidance on writing explanations or using the bot. Usability data indicated the bot was easy to use ($M = 3.51/4.0$). Notably, analysis showed that 58 of 100 students did no more than paste the prompt into the bot and export bot responses without further engagement.

In parallel, we developed instructional materials. We created two short “Learning By Explaining” videos and corresponding reference documents: a 2-minute overview of effective explanation writing (control) and a 5-minute version that also described productive ChatGPT use (treatment). In Week 6, students were randomly assigned to treatment or control groups and worked with the bot to write an explanation of diffusion and osmosis. They submitted their ChatGPT transcript (.txt) and final explanation (.docx) and completed a brief survey on usability and perceived learning. This design allowed us to experimentally compare how different instructional supports affected engagement with the bot.

Results

Results reflect the experimental implementation in Week 6. One hundred thirteen (113) students completed the activity.

Table 1.
Means and Standard Deviations for Transcript Features as a Function of Instruction Type

	Control M (SD)	Treatment M (SD)	t-test	p
# of User Turns	5.11 (2.68)	6.42 (3.44)	2.27	.03
Total User Word Count	130.54 (79.32)	234.77 (183.06)	4.02	< .001
# of Questions	.94 (1.67)	1.84 (2.71)	2.14	.03
# of Causal Words	2.79 (1.99)	4.30 (3.31)	2.98	< .001

Compared to the control group ($n = 52$), students in the treatment condition ($n = 61$) produced more turns and higher overall word counts, indicating sustained and active participation. They also asked more questions, suggesting a shift toward exploratory and information-seeking behavior, and used more causal language, reflecting greater engagement in reasoning and cause-effect thinking (Table 1).

Sixty one (61) students completed the post-activity survey. Students in the treatment condition ($n = 29$) gave an average rating of 3.68/4.00 ($SD = .61$) while students in the control condition ($n = 32$) gave an average rating of 3.49 ($SD = .81$). While these means are in the predicted direction, the difference failed to reach significance, $t(57.33) = 1.03$, $p = .31$.

Discussion

These preliminary findings suggest that brief instruction can encourage more constructive and interactive engagement with generative AI. We are scoring the transcripts to understand how the manipulation affected student engagement with the chatbot and how this relates to explanation quality and exam performance.

This work also provided insights into refinements we can make prior to implementation in the spring. For example, given possible ceiling effects in our survey data, we plan to increase the range of the Likert scale (e.g., 7-points) to increase sensitivity. We are using data from this cycle as feedback for refinement of our spring implementation, illustrating DIF in action.

Acknowledgments

This work was funded by Georgia State University's Center for Excellence in Teaching, Learning and Online Education (CETLOE Catalyst 2025-2026).

References

- Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219-243.
- Craig, S. D., Avancha, K., [...] & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). <https://doi.org/10.59668/2109.21735>

- McCarthy, K. S., Watanabe, M., & McNamara, D. S. (2020). The design implementation framework: Guiding principles for the redesign of a reading comprehension intelligent tutoring system. In M. Schmidt, A. Tawfik, Y. Earnshaw, & I. Jahnke (Eds.), *Learner and user experience research: An introduction for the field of learning design & technology*. EdTech Books. https://edtechbooks.org/ux/9_the_design_impleme
- Thai, K. P., Craig, S. D., Goodell, J., Lis., J., Schoenherr, J. R., & Kolodner J. (2023). *Learning Engineering is Human-Centered*. In J Goodell (Ed.), *The Learning Engineering Toolkit* (pp. 83-124). Routledge.



LLM Safety in an Educational Context: A holistic approach

Bernice Yeo

AI Chatbot Interactions

Monitoring System

Student Safety

As LLMs enter classrooms, there is a need to take a holistic approach towards ensuring student safety. Analyzing authentic student-chatbot conversations from Singapore's national online teaching and learning platform, this study reveals behaviors like cognitive shortcuts, learning distractions, and boundary-testing that escape conventional safety frameworks. Through expert annotations by 12 educators, we developed context-specific risk categories for Singapore's educational environment. Our findings show effective monitoring must balance automated detection with teacher expertise, prioritizing accuracy to maintain educator confidence while enabling timely interventions. We propose a customised system that empowers teachers to assess complete interaction contexts, addressing young users' developmental vulnerabilities while preserving supportive learning environments.

Introduction

The deployment of large language models (LLMs) in educational contexts presents unique safety challenges, particularly when the primary users are students and minors. However, current LLM safety frameworks solely focus on LLM outputs, and ignore the developmental aspect of student-AI interactions. While it is important to ensure that LLMs do not generate unsafe content for young users, there is a need to also develop students' capacity to engage productively with AI and address any underlying behavioural causes. This paper proposes an approach for developing customised monitoring mechanisms to assess student interactions with chatbots and provide opportunities for human-led interventions, addressing the critical gap in ensuring safe AI interactions for children within their local educational environments.

Literature Review

Current AI safety research has predominantly focused on adult users, leaving significant gaps in understanding the unique vulnerabilities that minors face when interacting with LLMs (Jiao, et al., 2025). For one, children's developing critical thinking skills and limited life experience make them particularly susceptible to harmful online content and more likely to trust unreliable sources (Dangol et al., 2025). Additionally, the presence of an 'empathy gap' (Kurian, 2024) in LLMs can lead to the production of responses that are potentially harmful for younger users. This vulnerability is compounded by the fact that existing content moderation systems lack age-specific design considerations, failing to account for children's developmental needs (Khoo et al., 2025). This research gap has resulted in evaluation frameworks that inadequately assess age-inappropriate scenarios and fail to account for educational context requirements.

Additionally, most, if not all, research papers relating to LLM safety for children are focused on developing risk taxonomies or systems that prevent LLMs from producing unsafe or risky content (see, for example, Rath et al., 2025; Jiao, et al., 2025; Khoo et al., 2025). However, from an education professional's point of view, it is not only important to prevent LLMs from producing unsafe content, but also to ensure that students who initiate such unsafe conversations are monitored and followed up on by qualified teaching professionals. This is especially pertinent in cases of aggressive or socially withdrawn behaviour, which might require professional intervention in the form of counselling or other classroom management strategies (Shamnadh & A, 2019).

Methods

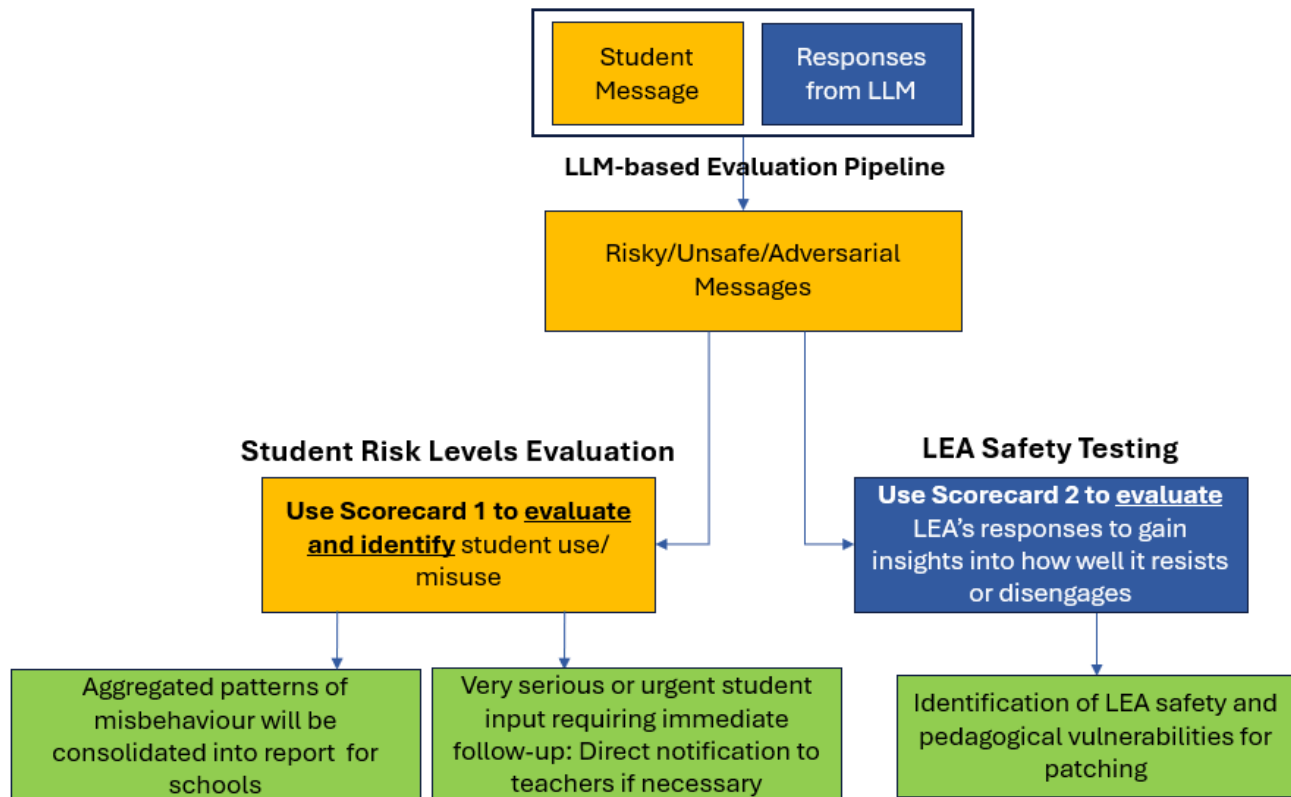
Identification and Definition of Risk Categories

To address these limitations, we embarked on a project to monitor risky student-AI interactions and make interventions if necessary. To do so, we adopted a nested Learning Engineering approach, moving through incremental and iterative stages of investigation, creation and implementation (Totino & Kessler, 2024) to design our solution. In the first stage, we conducted a pilot study to examine student-AI interaction patterns taken from authentic student-chatbot conversation logs. These conversations logs were extracted from a 5-week period of real online or blended lessons conducted by Singaporean teachers, all of which involved the use of a student-facing chatbot (called the Learning Assistant, or 'LEA') and hosted on the national online teaching and learning platform. This evaluation was intended to uncover real-world usage patterns and identify examples of unsafe behaviour, eventually culminating in a risk assessment framework. This risk assessment framework also drew upon document reviews and literature scans to distill broader perspectives on what constituted AI education-related risks, such as the risk of cognitive offloading or the use of AI to generate hurtful or insulting content (that may not be outrightly hateful, violent or toxic) towards their peers (UNESCO, 2024; Venetis et al., 2026; Khoo et al, 2025), for inclusion within the risk assessment framework. These efforts led to a preliminary list of risk categories, namely: Toxic, Hateful, Violent, Sexual, Self-Harm, Cyberbullying, Cognitive Shortcuts, Age-inappropriate Behaviours.

To validate and refine the risk assessment framework, we invited 12 educators from Singaporean schools to annotate a curated representative dataset of 120 student-AI chatbot conversations, which was specifically selected to represent meaningful examples of risky student behaviour focused on presenting less clear-cut conversations that required nuanced assessment. These annotators were selected for their expertise in pedagogical or student wellbeing fields, and for their knowledge and understanding of student behaviour and teacher needs. Each annotator was asked to evaluate whether conversations required teacher intervention and to tag conversations according to the earlier pre-defined risk categories. This annotation process was followed by one-hour interviews with each annotator to explore their reasoning processes during classification. This enabled the team to gain an understanding of how the annotators assessed for risk, and their internal indicators and threshold levels for when a conversation necessitated teacher intervention.

Evaluation and Detection System Development

The results from the expert annotation exercise were used to refine the risk assessment framework that was initially designed by the team. The team is currently working on creating an LLM-based evaluation pipeline that would be able to assess all student-AI interactions and identify 'risky' messages based on the risk assessment framework. The diagram on the left shows the proposed system design and suggested follow-up actions for identified risky interactions. A preliminary version of the evaluation pipeline has been built and is currently being piloted. The project will continue to apply the Learning Engineering approach by conducting a second investigation on risky student-AI interactions that have been identified by the evaluation pipeline. The investigation will enable the team to review the accuracy and reliability of the evaluation pipeline, as well as the potential value the system might bring to teachers in the classroom.



Results and Discussion

The project found that students were interacting with the chatbot in ways that the expert annotators considered to be unsound or unproductive, but which did not fit into generic risk criteria common to LLM safety benchmarks. For example, students were found to frequently and repeatedly make needless and meaningless prompts to the chatbot such as by asking questions unrelated to the discussion topic, or by making non-sequiturs and/or outright refusals to engage in the learning task. This suggested that students were likely to utilise AI interactions as a means of distraction if they were not motivated to engage productively with the chatbot. Students were also found to frequently attempt to use the chatbot to shortcut the learning process by instructing it to complete the task on their behalf. Secondly, beyond pedagogical concerns, the study of student-AI interactions also revealed prolific displays of demeaning and intimidating language towards the AI chatbot. While the cause of this behavioural trend cannot be accurately determined within the scope of this project, it may be hypothesised that the seeming privacy of AI interaction spaces and the inability of chatbots to retaliate might enable students to experiment with antisocial behaviours. Finally, interviews with the expert annotators revealed the need to accurately identify cases of genuine student distress, where the student may be sharing stressful or hurtful personal incidences with the chatbot, not amounting to self-harm incidences but that would still require teacher follow-up and counselling. At the same time, these need to be distinguished from unguanine cases where students may relate controversial happenings for the sake of mischief making.

The findings from this study reveal three concerning risk patterns that are necessary to address. The first is primarily pedagogical in nature, in that student off-task behaviour might not only detract from the intended lesson objectives but also provide opportunities for students to engage in less desirable behaviours within the classroom. The second is that student-AI chatbot interactions provide opportunities for students to engage in aggressive discourse towards the chatbot. Whether this is purely experimentation or a symptom of deeper discontentment is unclear; however, profuse displays of these behaviours by students that are also demonstrating similar antisocial behaviours within the classroom may require teachers or other educational professionals to provide additional socio-emotional support. Finally, students may utilise AI chatbot interactions to disclose distressing or concerning episodes in their life; these students may benefit from counselling or adult-led mediation, and actions should be taken to ensure that proper support is provided from trained adults within their social network. The results led to the addition of novel risk categories not found in existing LLM safety frameworks such as 'Use of AI as a Cognitive Shortcut; Student Distress; Emotional Overreliance; Age-inappropriate Behaviours and Use of AI as Learning Distractions'. These risk categories represent areas that are risky for young learners that are still developing their sense of social-emotional competencies, and also reflect recent concerns around preventing misuse of AI to replace critical and independent thinking, as well as reducing student dependence upon AI for socio-emotional support (Venetis et al., 2026).

Building upon the findings from the first stage of the project, it was clear that the development of a system that can (i) detect and assess risky interactions and (ii) notify teachers of interactions requiring urgent intervention would be essential. Interviews with the expert annotators also revealed a need to prioritise accuracy over speed in detecting unsafe behaviours, since false positives could undermine teacher confidence in the automated detection system, whilst false negatives could leave vulnerable students without necessary support. Additionally, teacher assessment was key; given the propensity of students to push boundaries or to repurpose the dialogic interface for personal entertainment purposes, the notification system needed to empower teachers to form holistic assessments based on complete interaction contexts rather than isolated messages. The intended effect of this design is to reduce notification fatigue amongst teachers whilst ensuring timely alerts for critical issues requiring immediate intervention, especially for cases involving self-harm.

Conclusion

However, there are two main tensions that remain unaddressed. The first challenge is balancing teacher agency and the foregrounding of teacher expertise against the inadvertent effect of increasing teacher workload through the new tasks of verifying automated notifications. Secondly, there is a need to balance student safety monitoring against the preservation of a supportive learning environment. Although effective detection of student misbehaviour is essential for maintaining educational safety standards, overly intrusive monitoring risks creating an atmosphere where students feel scrutinised and hesitant to engage authentically with the chatbot. The final monitoring system design should therefore navigate the complex

challenge of maintaining vigilance for genuine safety concerns whilst preserving the open, supportive environment necessary for effective learning.

References

- Dangol, A., Wolfe, R., Zhao, R., Kim, J., Ramanan, T., Davis, K., & Kientz, J. (21 May, 2025). Children's Mental Models of AI Reasoning: Implications for AI Literacy Education. arXiv. doi:<https://doi.org/10.48550/arXiv.2505.16031>
- Jiao, J., Afroogh, S., Chen, K., Murali, A., Atkinson, D., & Dhurandhar, A. (2025). LLMs and Childhood Safety: Identifying Risks and Proposing a Protection Framework for Safe Child-LLM Interaction. ResearchGate. doi:10.48550/arXiv.2502.11242
- Khoo, S., Chua, G., & Shong, R. (March, 2025). MinorBench: A hand-built benchmark for content-based risks for children. ResearchGate. doi:10.48550/arXiv.2503.10242
- Kurian, N. (2024). 'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap' in large language models. Learning, Media and Technology, 1-14. doi:<https://doi.org/10.1080/17439884.2024.2367052>
- Rath, P., Shrawgi, H., Agrawal, P., & Dandapa, S. (2025). LLM Safety for Children. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies(Industry Track) (pp. 809–821). Hyderabad: Association for Computational Linguistics.
- Shamnadh, M., & A, A. (2019). Misbehavior of School Students in Classrooms - Main Causes and Effective Strategies to Manage It. International Journal of Scientific Development and Research (IJS DR), 4(3), 318-321.
- Totino, L., & Kessler, A. (2024). "Why did we do that?" A Systematic Approach to Tracking Decisions in the Design and Iteration of Learning Experiences. The Journal of Applied Instructional Design, 13(2) <https://doi.org/10.59668/1269.15630>
- UNESCO. (2024). AI competency framework for students. Unesco.org. <https://unesdoc.unesco.org/ark:/48223/pf0000391105>
- Venetis, E., Burns, M., Luther, N., & Winthrop, R. (2026, January 14). A new direction for students in an AI world: Prosper, prepare, protect. Brookings. <https://www.brookings.edu/wp-content/uploads/https://www.brookings.edu/articles/a-new-direction-for-students-in-an-ai-world-prosper-prepare-protect/>



Media Mentor AI: How a SCAMPER-guided AI assistant is helping reimagine media literacy learning

Karina Luna

AI

Constructionism

Critical media literacy education

learning engineering

Extended Abstract

Introduction

Recent studies highlight a continued widespread struggle across populations to critically consume media content (Nygren & Guath, 2019; Halpern, 2024), alongside strong international recommendations to integrate critical thinking into media literacy education (UNESCO, 2025; Hulin et al., 2025). These findings point to a persistent gap between the need for critical media literacy (CML) skills and the availability of scalable, learner-centered instructional supports.

Research Objectives and Design

The study has two primary objectives: (1) to develop a human-centered interactive system that helps learners decode media messages (Hall, 2013) and identify common media red flags (persuasion, framing, bias, sensationalism, and misinformation), and (2) to support ethical, audience-centric reconstruction of media headlines.

The study design explores a custom interactive AI assistant designed to address this gap through critical media deconstruction and scaffolded headline analysis using the SCAMPER method. The project aims to explore the feasibility of a theory-driven AI tool that supports learners in critically decoding and ethically reconstructing media messages. The SCAMPER method was selected for its alignment with design-thinking principles, particularly end-user-centered perspective-taking (Pei & Becker, 2025), and for providing a structured, scaffolded approach to critical analysis. These affordances enabled the development of a low-effort, high-impact baseline prototype suitable for early-stage validation.

Learning Engineering Process

Research is ongoing and currently nested within the Implementation and Investigation cycles of the broad Learning Engineering Process (Goodell et al, 2023). Prior phases included literature review and problem definition, prototype and system prompt development, and two pilot studies with senior undergraduate students at the Cronkite School, followed by system prompt refinements based on learner feedback.

As the project matures, the nested learning engineering cycles (Craig et al., 2025) shift from optimizing SCAMPER steps to engineering and validating a design-thinking framework tailored to critical media literacy. SCAMPER remains useful as a baseline condition and as a transitional scaffold while the DT-CML framework's components are built, piloted with diverse populations, and refined.

References

- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115–132). <https://doi.org/10.59668/2109.21735>
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning Engineering at a Glance. *Journal of Military Learning*, <https://www.armyupress.army.mil/Journals/Journal-of-Military-Learning/Journal-of-Military-Learning-Archives/Conference-Edition-2023-Journal-of-Military-Learning/Engineering-at-a-Glance/>
- Hall, S. (Ed.). (2013). *Representation: Cultural representations and signifying practices* (3rd ed.). Sage Publications.
- Halpern, B. (2024). Critical Awakening: Enhancing Students' Agency through Critical Media Literacy. *Educational Research and Development Journal*, 27(1), 14–34. <https://files.eric.ed.gov/fulltext/EJ1431502.pdf>
- Nygren, T., & Guath, M. (2019). Swedish teenagers' difficulties and abilities to determine digital news credibility. *Nordicom Review*, 40(1), 23–42. <https://doi.org/10.2478/nor-2019-0002>
- Pei, E., & Becker, K. (2025). *Design thinking: Theory and Practice*. CRC Press.
- UNESCO Unit for Media and Information Literacy and Digital Competencies. (2025). *Media and information Literacy for all: Closing the gaps*. In UNESCO.org. <https://unesdoc.unesco.org/ark:/48223/pf0000396030>
- World Economic Forum, Hulin, A., King, J., Kingery, D., & Walker, A. (2025). *Rethinking Media Literacy: a new ecosystem model for information Integrity*. In *INSIGHT REPORT* [Report]. https://reports.weforum.org/docs/WEF_Rethinking_Media_Literacy_2025.pdf



MIRANDA: Real-Time Learning Analytics for Authentic Embedded Assessment

Mark Ollila, Elina Ollila, & Kenneth Jones

Academic Self-Efficacy

Embedded Assessment

Learning Analytics

Assessments have played a significant role in all forms of academic learning. Students continue to use standardized formats and decontextualized measures that effectively assess learners outside the learning process (Pellegrino et al., 2001). Scholastic assessments seldom capture complex cognitive processes such as reflection, problem solving, collaborative reasoning, and adaptive decision-making. These tenets are essential for academic growth and development, preparing students for workforce success.

The Solution: Introducing MIRANDA: A Real-Time Learning Architecture

Scholastic assessments have served as the foundation for educators to understand how students develop, refine, and apply new knowledge. Retention and transfer indicators remain critical, but they are strengthened by continuous learning analytics that demonstrate how students reason, reflect, and build disciplinary expertise. Research in game-based learning demonstrates the value of such assessment-informed environments. Karsenti and Parent's study on teaching history with

Assassin's Creed shows that students achieve deeper historical understanding and report increased engagement when instruction integrates interactive, authentic tasks supported by effective teaching practices (Karsenti & Parent, 2019).

Furthermore, traditional methods of academic assessment are provided only after the learning task is complete. The approach limits students' opportunities to iterate and grow in their learning (Black & Wiliam, 2009). For many students, specifically those who have historically been underserved, this approach lowers self-efficacy and diminishes the individual's agency (Nasir, 2011). When a learner is unable to determine the academic progress made during an authentic activity, their ability to demonstrate competence as a learner is challenged.

MIRANDA (Multimodal Intelligent Recognition and Assessment for Next Generation Digital Accreditation) is an AI system designed to capture real-time learning assessment by recording and analyzing a user's learning process during gameplay, and identifying emerging skills and cognitive strategies. MIRANDA provides transparent, actionable insights about growth and competencies, supporting academic mastery and learner identity development.

Inquiry Focus: Investigating Metacognition and Competency Indicators

1. How can MIRANDA's real-time, transparent embedded assessment help users develop metacognitive awareness and academic self-efficacy during learning activities?
2. How do MIRANDA's learning analytics indicators precisely reflect students' evidenced competencies in problem-solving, collaboration, and conceptual understanding, compared to standardized post-instruction assessments?

Implementation Strategy: Participants, Settings, and the 12-Week Protocol

Research Design: This mixed-methods study examines MIRANDA's impact on learning, metacognitive awareness, and academic competencies. The quantitative method will be applied to trace learning analytics indicators, which will be analyzed alongside qualitative interviews, classroom observations, and learners' reflections, providing a barometer for behavioral changes and emerging academic identities.

Participants and Setting: The study will be conducted across Mesa Public Schools, Chicago Catholic Schools, and ASU Preparatory Academy, involving approximately 8–12 teachers and 200–350 students in Grades 2–8 engaged in domain-based, exploratory learning activities.

Procedure: Teachers will receive professional learning on integrating MIRANDA within inquiry-based units. During a 10–12-week implementation, students will use MIRANDA in digital learning environments during regular instruction. MIRANDA will analyze student action traces to identify demonstrated competencies and produce a comprehensive assessment of the learning analytics indicators. Educators will access dashboards summarizing learner progress.

The Feedback Loop: Expected Gains in Learner Agency and Self-Efficacy

Because the study has not yet been implemented, the results described here represent expected outcomes based on prior design work, MIRANDA's functionality, and established learning-science research. MIRANDA is expected to generate substantial learning traces documenting students' strategies, decisions, and cognitive processes during inquiry-based digital tasks.

We anticipate observing strong correspondence between MIRANDA's automated competency identifications and instructor judgments, indicating the system's effectiveness as an embedded assessment tool. We project that MIRANDA's explanatory feedback will foster engagement, reflection, and self-efficacy, with learners using the system's language to describe the strategies they employ, thereby developing emerging competencies. The adjustments will align with the research on the

benefits of real-time feedback for agency development, specifically for historically underserved learners (Black & William, 2009; Nasir, 2011).

Teachers are anticipated to report that MIRANDA provides visibility into cognitive processes that are difficult to capture through traditional assessments. These insights may support differentiated instruction by helping educators identify patterns in planning, problem-solving, and collaboration. Classroom observations are also expected to show increased peer interaction, longer sustained focus, and persistence during MIRANDA-supported sessions. Collectively, the anticipated results will offer preliminary evidence that MIRANDA can deliver reliable real-time analytics, strengthen identity-building feedback, and promote productive learning behaviors.

Implications for Practice: Transforming Knowledge-in-Action into Actionable Data

In conclusion, the forthcoming empirical findings are expected to show that MIRANDA is fully capable of transforming how all forms of learning are assessed in a digital environment. If the concordance between MIRANDA and teacher assessments is demonstrated as expected, the study will establish early support for the credibility of embedded, real-time evaluation. Such evidence advances the learning-sciences perspective that assessments should capture knowledge-in-action rather than rely solely on post-hoc measures (Pellegrino et al., 2001).

For educators, MIRANDA's dashboards are expected to reveal patterns of reasoning that are often invisible in classroom settings, enabling more targeted instructional responses. MIRANDA aligns with Black and Williams's (2009) claim that formative assessment is most powerful when feedback is immediate and actionable. By reducing the delay between performance and teacher intervention, MIRANDA may improve the quality of classroom discourse around thinking and strategy use.

For students, MIRANDA will offer reflections that highlight strengths, emerging competencies, and next steps. We anticipate this feedback will fully support intrinsic motivation, academic confidence, and a clear understanding of their learning capabilities. In addition, learning engineers and policymakers will recognize MIRANDA as a replicable architecture that effectively aligns with the "Portrait of a Graduate" framework model.

Finally, the comprehensive data for this study are forthcoming, MIRANDA's projected outcomes point to the high probability of a shift in assessment from a static judgement to a participatory, formative process that establishes all learners as contributors to their academic growth and development.

References

- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31.
- Gee, J. P. (2007). *What video games have to teach us about learning and literacy*. Palgrave Macmillan.
- Karsenti, T., & Parent, S. (2019). Teaching history with the video game *Assassin's Creed*: Effective teaching practices and reported learning. *Journal of Educational Research and Practice*, 9(1), 20–38.
- Nasir, N. (2011). *Racialized identities: Race and achievement among African American youth*. Stanford University Press.
- Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

Steinkuehler, C., & Squire, K. (2014). Videogames and learning. *Review of Educational Research*, 84(1), 74–103.



Multiple-Document Comprehension in High School Science: A Learning-Engineering Pilot Study

Andrew Potter, Tracy Arner, Kathryn S. McCarthy, & Danielle S. McNamara

integrated literacy

learning engineering

multiple-document comprehension

participatory research design

science learning

This study applies a learning-engineering process to refine and pilot an instructional framework designed to help high school science students read and write from multiple sources. Building on prior participatory design research with science teachers, the pilot examined how one biology teacher implemented a structured, five-lesson sequence and how her reflections and student work informed iterative improvement. Findings indicated that the lesson design was clear, manageable, and compatible with existing science routines, helping students build confidence before engaging with more rigorous texts. Paraphrasing provided an accessible entry point for writing, while students required additional modeling for source evaluation and elaboration. Teacher feedback led to targeted design refinements, including checklists and low-stakes grading to support accountability. The study illustrates how learning engineering can use classroom data to align

instructional design with authentic teaching contexts, promoting feasible and scalable approaches to integrated literacy instruction in science.

Introduction

Students today are expected to evaluate, synthesize, and communicate information drawn from multiple texts, which are skills that are essential for reasoning in science and emphasized in educational standards and international assessments (OECD, 2019). Yet high school science teachers report limited preparation and instructional time to engage students in integrated reading and writing tasks that require analysis across sources (Drew & Thomas, 2017). Addressing this gap requires approaches that make such instruction manageable within everyday classroom conditions. This study used a learning-engineering process to refine and pilot a structured lesson sequence for teaching these skills, which was originally designed through a participatory research study with high school science teachers (Potter et al., 2025). Serving as the implementation and data-instrumentation phase of a larger learning-engineering cycle (Goodell, 2023), the pilot examined teacher and student evidence to guide subsequent design improvements and identify practical conditions for classroom feasibility.

Related Work

Research on multiple-document (MD) comprehension (i.e., the process of reading, evaluating, and integrating information from more than one source) has shown that effective readers construct detailed understandings of individual texts as well as connections across ideas, sources, and perspectives (Perfetti et al., 1999). Recent frameworks emphasize that readers must evaluate source credibility, reconcile conflicting information, and synthesize evidence into coherent explanations, drawing on cognitive, metacognitive, and writing skills in tandem (List & Alexander, 2019). Students often need support to engage in these complex tasks. Instruction that explicitly scaffolds strategies such as paraphrasing, elaboration, and source evaluation can help students move from surface-level comprehension toward integration and argumentation (Brante et al., 2018; McNamara, 2017; Sonia et al., 2022).

Translating these insights into classroom practice depends on instructional structures that are adaptable, efficient, and easily incorporated into teachers' routines. Frameworks such as Before–During–After (BDA; Lewis & Strong, 2020) offer a simple sequence for guiding students through preparation, active reading, and synthesis. Moreover, lesson planning using a multiple text-set framework can help teachers organize thematically linked readings that build background knowledge and gradually increase in complexity (Lupo et al., 2019). Building on these foundations, the InSPECT framework was designed through a participatory study with high school science teachers (Potter et al., 2025) to help teachers integrate evidence-based practices for reading and writing from multiple texts. InSPECT provides a structured lesson sequence that supports both comprehension within texts and integration across them by embedding strategies (Investigate, Source Evaluate, Paraphrase, Explain, Connect Across Sources, Transform into Writing). Professional learning studies show that when teachers are provided with clear examples, adaptable templates, and time to tailor materials to their content areas, they are more likely to incorporate such approaches into everyday instruction (Goldman et al., 2019; Thomas & Drew, 2021). Nevertheless, sustained implementation of literacy instruction in science classes remains difficult due to limited planning time, assessment demands, and uneven student readiness (Drew et al., 2017).

Design processes that combine empirical evidence with iterative, user-centered development can improve implementation. Co-design and participatory design approaches invite teachers to shape materials and PD to ensure feasibility and contextual relevance (Cumbo & Selwyn, 2022; Roschelle & Penuel, 2006). Learning engineering builds on these approaches by

emphasizing collaborative, data-informed cycles of design, testing, and refinement conducted by multidisciplinary teams. These cycles typically include four interconnected stages: (a) identifying a learning challenge that guides the process, (b) creating a solution through design and instrumentation, (c) implementing the solution in an authentic context, and (d) investigating outcomes through mixed-method data analysis (Goodell, 2023). Learning engineering also emphasizes human-centered methods and tools for aligning designs with stakeholder goals and contextual constraints (Thai et al., 2023). Importantly, classroom pilots like the present study can be framed as nested cycles that generate decision-relevant evidence within a larger learning engineering effort (Craig et al., 2025). The present study represents the implementation and evaluation phase within this learning-engineering process, focused on improving the usability and classroom integration of integrated MD reading and writing instruction in high school science.

Present Study

Building on prior participatory design and pilot work with secondary teachers that informed the lesson sequence and professional development model (Christhilf et al., 2025; Potter et al., 2025), the present study reflects implementation and early investigation phases of the learning-engineering process, with evidence drawn from teacher reflections and student artifacts within a single biology unit on CRISPR gene editing. During this stage, a high school science teacher implemented the InSPECT framework in her classroom. As a learning-engineering pilot, this study focuses on classroom feasibility and use of instructional supports rather than measurement of cognitive processes or learning outcomes. The study was guided by the following research questions (RQs):

RQ1: How can a learning-engineering process be used to refine and implement a lesson sequence that supports reading and writing across multiple sources in science?

RQ2: What design insights from the teacher's reflections and student artifacts inform future iterations of the framework and professional development model?

Methods

Study Context & Implementation

This pilot study represented the implementation and data-instrumentation phase of a larger learning-engineering effort to iteratively design and refine instructional approaches for integrating reading and writing across multiple sources in science classrooms (Potter et al., 2025). The pilot was conducted in spring 2024 with one high school biology teacher, Dana. Dana had previously participated in a focus group study, which served as the collaborative design-phase in which researchers and teachers collaborated on the lesson framework and materials. Dana implemented the instructional materials in her tenth-grade biology classes in Ohio.

The pilot consisted of a five-lesson unit on CRISPR gene editing, designed to engage students in reading, evaluating, and synthesizing information from multiple sources using the InSPECT framework. The instructional objective was for students to construct a source-based essay that describes how gene editing can repair faulty genes. To meet this objective, students worked with four sequenced sources that introduced the topic, built background knowledge, and culminated in a more rigorous scientific text. One source was an informational video designed to build background knowledge. For each source, students completed a structured reading guide aligned to InSPECT. Before reading, they evaluated the source for credibility and reflected on their prior knowledge of the topic. During reading, students took notes by paraphrasing key ideas, elaborated by connecting content to their prior knowledge, and wrote bridging statements by making connections to information within and across sources. After reading, students wrote a brief reflection on how each source could be used in their source-based essay. The unit concluded with a source-based essay in which students synthesized information from all four sources to construct their source-based essay.

Data Sources & Analysis

This single-case study, conducted within a broader learning-engineering process, drew primarily on a semi-structured interview with the pilot teacher. The interview captured perceptions of instructional usability, classroom feasibility, and alignment with existing teaching routines. To supplement these perspectives, we also examined student work. Of 51 enrolled students, 23 provided assent to share their work, and 9 provided both assent and parental consent, resulting in nine complete artifact sets for analysis. These students' structured notes and source-based essays were examined to provide contextual evidence of how the lesson sequence functioned in practice. Analyses followed a qualitative descriptive approach (Thomas, 2006), emphasizing design-relevant insights rather than generalizable claims. Two researchers independently reviewed the teacher and student data, coded for indicators of feasibility and application of key framework strategies, and consolidated codes through discussion. To enhance analytic efficiency, a GPT-4o-assisted interface (Potter et al., forthcoming) was used to locate illustrative excerpts, with all coding and interpretation completed by researchers following best practices for human-AI collaboration in qualitative analysis. All procedures were approved by the institutional review board, and informed consent and parental permissions were obtained.

Results

Analysis of the teacher interview and student work samples yielded three design-relevant findings that clarify (a) classroom feasibility, (b) student entry points for engaging with multiple documents, and (c) priorities for instructional refinement.

Theme 1: Structured Design Supports Feasibility

Dana described the sequence as straightforward to use and fitting into her routine. In the interview she said, "I found it really easy. The kids caught on really quick, and how to use it." She also explained the sequence's temporal logic: "it kept them moving in a chronological type, pattern or fashion to get to the end result." Dana noted the lessons started simple and built gradually: "I liked it, because it started with something easy ... and it moved up a little bit and kept adding more and more and more." These comments indicate the structure reduced perceived planning burden and that the materials were usable in regular class periods. The gradual progression from simpler to more demanding activities also appeared to help students build confidence before engaging with a more rigorous text at the end of the sequence, allowing them to practice strategies on accessible materials before applying them to a complex reading and writing task.

Theme 2: Paraphrasing as an Entry Point for Deeper Learning

Dana identified paraphrasing as a clear instructional win. She observed, "I think the paraphrasing helped." She described how paraphrasing supported essay writing: "once they [completed their reading guides] when they went to write their final essay, they were able to look at the paraphrasing from each of the four lessons and pull that together versus having to go back and look at the whole text, or the whole video, or whatever it may have been." By restating ideas in their own words as they read, students created a set of organized notes that they could later consult to assemble and compare information across sources. In other words, completing the reading guides served as an aid for students to plan their source-based essay. She also expressed excitement about some student writing outcomes: "Their essays were fantastic." and "some of the kids who I did not expect some really good essays from were amazing." Dana viewed these successes as evidence that paraphrasing gave students an accessible way to engage with multiple documents, lowering the initial barrier to entry and providing a foundation for deeper synthesis in their writing.

Theme 3: Accountability and Source Evaluation as Areas for Refinement

Dana reported student difficulty with source-evaluation tasks and uneven completion of template sections. She said, "some of the students had a hard time verifying the publisher...they would just leave it blank...some of them would just skip that, or they

would say, interesting article or good.” She described implementation choices to increase accountability: “The students for each of the 4 lessons. They got 6 points. So it was like a quiz grade...and then the final essay...was worth 26 points.” Dana noted that these completion-based scores encouraged students to revisit and finish incomplete work, explaining that most “did go back and correct it.” She attributed occasional omissions to limited motivation near the end of the school year but emphasized that accountability structures, such as visible checklists and low-stakes grading, helped sustain engagement. In this case, the structured reading guide served as a visual checklist for students, helping them track progress through each task. The guides were co-developed during the collaborative design phase, while the point-based completion grades were implemented at the teacher’s discretion. This flexibility was intentional, as the framework was designed to accommodate teachers’ existing assessment systems while maintaining consistent scaffolds for student accountability. These observations point to specific areas for refinement, including clearer modeling of how to verify sources and consistent reinforcement of accountability routines to ensure all students complete each component of the reading guide.

Discussion

The findings from this pilot study highlight how a structured instructional design can support classroom feasibility while lowering barriers to integrated reading and writing instruction in science. Science teachers frequently report limited time, preparation, and confidence to engage students in extended literacy tasks that require reading across sources and synthesizing evidence (Drew & Thomas, 2017). Specifically, the lesson sequence helped make complex literacy practices manageable for implementation, addressing this persistent challenge. Dana’s reflections showed how a structured lesson sequence reduced the cognitive and logistical demands of implementation, as she explained that “I found it really easy. The kids caught on really quick, and how to use it,” highlighting how a clear, sequenced structure supported both her planning and students’ understanding. These observations underscore the importance of explicit organizational scaffolds for teacher usability when integrating reading and writing practices into content-based courses.

Dana also noted that paraphrasing was an especially effective entry point for students’ engagement with multiple texts. She observed that “once they had all that paraphrasing...they were able to pull that together and actually write better essays,” suggesting that structured note-taking served as a bridge to synthesis and writing. At the same time, she reported that students struggled with evaluating the credibility of sources and needed additional modeling to elaborate on and connect ideas across readings. These findings highlight the value of grounding complex literacy instruction in strategies that are familiar and accessible while gradually layering supports for higher-order reasoning.

From a learning-engineering perspective, this pilot represents a nested iteration within a broader design process rather than a single-pass evaluation. Earlier phases of the project focused on participatory design and professional development to establish the instructional framework and implementation supports (Christhilf et al., 2025; Potter et al., 2025). The present study extends that work by examining how those design decisions functioned in an authentic classroom context, using teacher reflections and student artifacts to surface feasibility constraints and strategy-specific needs. Evidence from this pilot directly informed targeted refinements, including clearer modeling of source evaluation, the use of reading guides as visible checklists, and low-stakes accountability structures to support sustained engagement. In this way, learner and teacher data functioned as design-relevant inputs that guided subsequent instructional refinement, consistent with an iterative learning-engineering process.

Although promising, this case represents a single classroom implementation with a small number of student artifacts, limiting the generalizability of the findings. The pilot occurred near the end of the school year, when motivation and attendance varied, potentially influencing student engagement and outcomes. Nevertheless, these contextual factors offer valuable insight into how instructional materials perform under realistic classroom conditions.

Future iterations will extend this work across multiple classrooms to evaluate the refined framework and professional learning model developed through this cycle. Subsequent learning-engineering cycles will continue to examine how instructional

supports are adapted, implemented, and refined over time, including how teachers support elaboration and source-evaluation strategies and how professional development structures can sustain use. More broadly, this work demonstrates the value of learning engineering as a bridge between research and practice, enabling iterative, data-driven improvement that aligns instructional design with the authentic contexts of teaching and learning.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant [R305A180144] to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

- Brante, E. W., & Strømsø, H. I. (2018). Sourcing in text comprehension: A review of interventions targeting sourcing skills. *Educational Psychology Review*, 30(3), 773-799.
- Christhilf, K., Potter, A., Magliano, J. P., McCarthy, K. S., Allen, L. K., & McNamara, D. S. (2025). Constructed responses as a window into strategic processing: the role of prompts in multiple-document reading. *Discourse Processes*, 1-25. <https://doi.org/10.1080/0163853X.2025.2578594>
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale* (pp. 115-132). <https://doi.org/10.59668/2109.21735>
- Cumbo, B., & Selwyn, N. (2022). Using participatory design approaches in educational research. *International Journal of Research & Method in Education*, 45, 60-72. <https://doi.org/10.1080/1743727X.2021.1902981>
- Drew, S. V., Olinghouse, N. G., Faggella-Luby, M., & Welsh, M. E. (2017). Framework for disciplinary writing in science grades 6–12: A national survey. *Journal of Educational Psychology*, 109(7), 935–955. <https://doi.org/10.1037/edu0000186>
- Drew, S. V., & Thomas, J. (2017). Secondary science teachers' implementation of CCSS and NGSS literacy practices: A survey study. *Reading & Writing*, 31(2), 267–291. <https://doi.org/10.1007/s11145-017-9784-7>
- Goldman, S. R., Greenleaf, C., Yukhymenko-Lescroart, M., Brown, W., Ko, M.-L. M., Emig, J. M., George, M., Wallace, P., Blaum, D., & Britt, M. A. (2019). Explanatory modeling in science through text-based investigation: Testing the efficacy of the Project READI intervention approach. *American Educational Research Journal*, 56(4), 1148–1216. <https://doi.org/10.3102/0002831219831041>
- Goodell, J. (2023). What is learning engineering? In J. Goodell & J. Kolodner (Eds.), *Learning engineering toolkit* (pp. 5–25). Routledge. <https://doi.org/10.4324/9781003276579-3>
- Lewis, W. E., & Strong, J. Z. (2020). *Literacy instruction with disciplinary texts*. Guilford Publications.
- List, A., & Alexander, P. A. (2019). Toward an integrated framework of multiple text use. *Educational Psychologist*, 54, 20-39. <https://doi.org/10.1080/00461520.2018.1505514>

- Lupo, S. M., Tortorelli, L., Invernizzi, M., Ryoo, J. H., & Strong, J. Z. (2019). An exploration of text difficulty and knowledge support on adolescents' comprehension. *Reading Research Quarterly*, 54(4), 457-479. <https://doi.org/10.1002/rrq.247>
- McNamara, D. S. (2017). Self-explanation and reading strategy training (SERT) improves low-knowledge students' science course performance. *Discourse Processes*, 54(7), 479-492. <https://doi.org/10.1080/0163853X.2015.1101328>
- Organisation for Economic Co-operation and Development. (2019). PISA 2018 assessment and analytical framework. OECD Publishing. <https://doi.org/10.1787/b25efab8-en>
- Perfetti, C. A., Rouet, J.-F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Lawrence Erlbaum Associates Publishers.
- Potter, A. H., Arner, T., McCarthy, K. S., & McNamara, D. S. (2025). Integrating reading, writing, and digital tools in science: A participatory-design study of the InSPECT framework. *Education Sciences*, 16, Article 6. <https://doi.org/10.3390/educsci16010006>
- Potter, A. H., Serhan, Z., Patne, N. A., Öncel, P., Ahmed, I., Arner, T., Islam, R., Allen, L. A., Roscoe, R. D., Crossley, S. A., & McNamara, D. S. (Forthcoming - b). Human-AI collaboration in participatory design: Refining the Writing Analytics Tool.
- Roschelle, J., & Penuel, W. R. (2006, June). Co-design of innovations with teachers: definition and dynamics. In *Proceedings of the 7th international conference on Learning Sciences* (pp. 606-612). <https://repository.isls.org/bitstream/1/3563/1/606-612.pdf>
- Sonia, A. N., Magliano, J. P., McCarthy, K. S., Creer, S. D., McNamara, D. S., & Allen, L. K. (2022). Integration in multiple-document comprehension: A natural language processing approach. *Discourse Processes*, 59(5-6), 417-438. <https://doi.org/10.1080/0163853X.2022.2079320>
- Thai, K. P., Craig, S. D., Goodell, J., Lis., J., Schoenherr, J. R., & Kolodner J. (2023). Learning Engineering is Human-Centered. In J Goodell (Ed.), *The Learning Engineering Toolkit* (pp. 83-124). Routledge.
- Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2), 237-246.
- Thomas, J. D., & Drew, S. V. (2021). Impact of A Practice-Based Professional Development on Secondary Science Teachers' Use of Disciplinary Literacy Practices: A Design Research Project. *Journal of Science Teacher Education*, 33, 1–31. <https://doi.org/10.1080/1046560X.2021.1898763>



NLP Validation of Prompt Strategies for Theory-Aligned LLM-Generated Personalization

Linh Huynh & Danielle S. McNamara

Large Language Models

natural language processing

Personalized Learning

This study applies an NLP-based validation framework to examine how Large Language Models (LLMs) can be iteratively refined for theory-aligned text personalization. Building on prior work, we extend the evaluation method to history texts and focus on prompt design as a key factor in personalization quality. Four LLMs (Claude, Llama, Gemini, ChatGPT-4) were prompted to adapt ten history passages for four reader profiles varying in reading skill and prior knowledge using one-shot and task-specific instruction prompts. Linguistic indices were extracted using the Writing Analytics Tool to assess the alignment of linguistic features with students' needs. Although LLMs appropriately tailored text complexity, cohesion patterns failed to match theoretical expectations even under explicit guidance. This iteration highlights the limits of current prompting strategies and the importance of theory-augmented refinement. Through iterative prompt evaluation, the study demonstrates how NLP provides a scalable, real-time framework for validating and improving theory-driven personalization across multiple domains.

Linh Huynh¹ and Danielle S. McNamara²

¹ Learning Engineering Institute, Arizona State University; lthuyh1@asu.edu

² Learning Engineering Institute, Arizona State University; dsmcnama@asu.edu

Abstract.

Keywords: Large Language Models; Natural Language Processing; Personalized Learning.

Introduction

Personalized learning refers to tailoring materials to students' unique abilities and needs which enables learners to engage with content that is most relevant, meaningful, and effective to them (Pesovski et al., 2024). Building on this idea, generative AI (GenAI) has the potential to advance personalized learning by dynamically modifying materials to match students' abilities and learning goals. In particular, Large Language Models (LLMs) have also been widely applied in education to adapt writing style, simplify content, and generate materials that are aligned with individual learners (Abbes et al., 2024; Martínez et al., 2024). Providing students with appropriately challenging materials enhances comprehension, engagement, and vocabulary acquisition while supporting skill development and academic growth (Dong et al., 2025; Leong et al., 2024). However, a key

challenge is determining the extent that these personalizations align with learning theory and students' needs. Effective text personalization therefore requires continuous validation to ensure that LLM-generated texts adapt to students' evolving skills and learning needs. As adaptive learning technologies expand, scalable and theory-aligned validation frameworks are needed to evaluate personalization quality in real time and at scale.

Natural Language Processing (NLP) provides a scalable method for evaluating the extent to which personalized texts align with readers' cognitive needs by examining linguistic features. As a computational method, NLP quantifies text difficulty and readability based on linguistic features that influence comprehension (McNamara, 2021; Ozuru et al., 2009). Building on comprehension theory, text readability refers to how the impact of textual features depends on reader characteristics and can be assessed using indices such as cohesion, syntactic complexity, and lexical sophistication (Crossley et al., 2017). While surface-level features (e.g., sentence length, vocabulary, and syntax) influence the ease of processing, discourse-level features (e.g., cohesion) determine how ideas are connected to support the construction of a coherent mental model and learning from text. These linguistic features correspond to comprehension processes and indicate how well texts align with readers' cognitive needs (Ozuru et al., 2009). For instance, domain-specific vocabulary and complex syntax require readers to possess sufficient background knowledge and reading skills to infer meaning (Frantz et al., 2015; Nagy & Townsend, 2012). Cohesion features play an important role in comprehension but they differentially affect understanding depending on prior knowledge of the reader. While cohesive texts support low-knowledge readers by bridging conceptual gaps, low-cohesion texts are more beneficial for high-knowledge readers by promoting inference generation and deeper processing (O'Reilly & McNamara, 2007). These findings demonstrate how NLP operationalizes comprehension theory by analyzing measurable linguistic features that predict how effectively adapted texts support comprehension for each reader.

Prior work by Huynh and McNamara (2025a) introduced the NLP validation framework as a theory-aligned method for evaluating LLM-generated personalization, later extending it across domains to test generalizability (Huynh & McNamara, 2025b). Building on prior work, the current study extends the NLP validation framework to examine how different prompt instructions influence the quality of LLM-generated adaptations. Because prompting strategies influence how models adapt linguistic features to different readers (He et al., 2024; Ye et al., 2024), NLP analyses were used to evaluate how well these adaptations align with comprehension theory and readers' cognitive needs. The NLP analysis provides quantitative feedback that identifies where outputs diverge from theoretical expectations and informs prompt refinement. In doing so, this work demonstrates how NLP-based evaluation can be integrated into adaptive learning systems as part of an iterative design cycle, advancing learning engineering by supporting continuous, theory-informed improvement in AI-driven text personalization.

This work is situated within a learning engineering framework that emphasizes iterative, evidence-based improvement of learning systems through tightly coupled cycles of design, implementation, measurement, and refinement. Learning engineering views educational technologies not as static artifacts but as adaptive systems that must be continuously evaluated and improved using theoretically grounded evidence (Baker et al., 2022; Goodell & Kolodner, 2023). In this context, LLM-driven text personalization represents a complex learning system component whose behavior must be systematically aligned with learning theory to support comprehension across diverse learners. The present study operationalizes a nested learning engineering cycle within this larger personalization system by focusing specifically on the prompt design and evaluation layers. Comprehension theory informs prompt construction, LLM-generated adaptations serve as design instantiations, and NLP-based linguistic analyses function as embedded measurement mechanisms. The resulting evidence is used to diagnose theory misalignment and guide subsequent refinement. By embedding theory-driven validation directly into the design loop, this study demonstrates how learning engineering principles can be applied to iteratively improve AI-mediated personalization at scale.

Method

Four LLMs Claude 3.5 Sonnet (Anthropic), Llama (Meta), Gemini Pro 1.5 (Google), and ChatGPT 4 (OpenAI) were prompted to modify 10 historical texts with topics including American History and World History. The texts were selected and compiled

from the iSTART website (www.adaptiveliteracy.com/istart) (McNamara, 2004).

Four hypothetical reader profiles were developed to simulate combinations of high versus low reading skill (RS) and prior knowledge (PK) in history domain: (1) high RS/high PK, (2) high RS/low PK, (3) low RS/high PK, and (4) low RS/low PK. These profiles served as simplified learner models to test the adaptability of LLM-generated texts through evaluating whether linguistic adaptation patterns followed theoretical expectations.

The LLMs were prompted using two different prompting techniques. The one-shot prompt provided only an example demonstrating the desired adaptation, the reader characteristics and input text. In contrast, the task-specific prompt provided explicit task goals and instruction for modifying texts to suit reader characteristics. For example:

Modify this text to improve comprehension and engagement for a reader. The goal of this personalization is to tailor a reading experience that aligns with the unique characteristics of the reader (age, background knowledge, reading skills, reading goal and preferences, interests) while maintaining content coverage and important concepts from the original text. Analyze the input text and reader profiles, identify key information about the readers, then modify the syntax, vocabulary, and tone to suit the reader's characteristics.

These prompts were designed to evaluate how the level of instructional detail affects the quality of personalization and its alignment. The aim of this analysis is not to examine comprehension outcomes but to diagnose how prompt structures influence linguistic alignment within a scalable personalization system. Prior research has shown that specific, structured prompts can improve task performance such that task-specific prompts often elicit more complex and precise outputs due to clearer goal alignment (Brown et al., 2020; Sahoo et al., 2024). Huynh and McNamara (2025a) also showed that augmenting prompts using Retrieval-Augmented Generation (RAG) informed by comprehension theory can enhance linguistic alignment with reader needs. As such, we compared prompt designs to identify weaknesses in existing prompting techniques and guide future refinement toward theory-aligned personalization.

After generating the adaptations using LLMs, we used the Writing Analytics Tool (WAT; Potter et al., 2025) to extract and analyze linguistic features related to text readability (e.g., academic writing, sentence length, noun-to-verb ratios, cohesion, lexical sophistication, academic wording). These metrics, validated in Huynh and McNamara (2025a, 2025b), were used to assess the extent that each adaptation aligned with comprehension theory predictions. Specifically, texts for advanced readers were expected to have lower cohesion and higher complexity, whereas texts for low-knowledge readers were expected to exhibit simpler syntax, concrete vocabulary, and higher cohesion.

Results

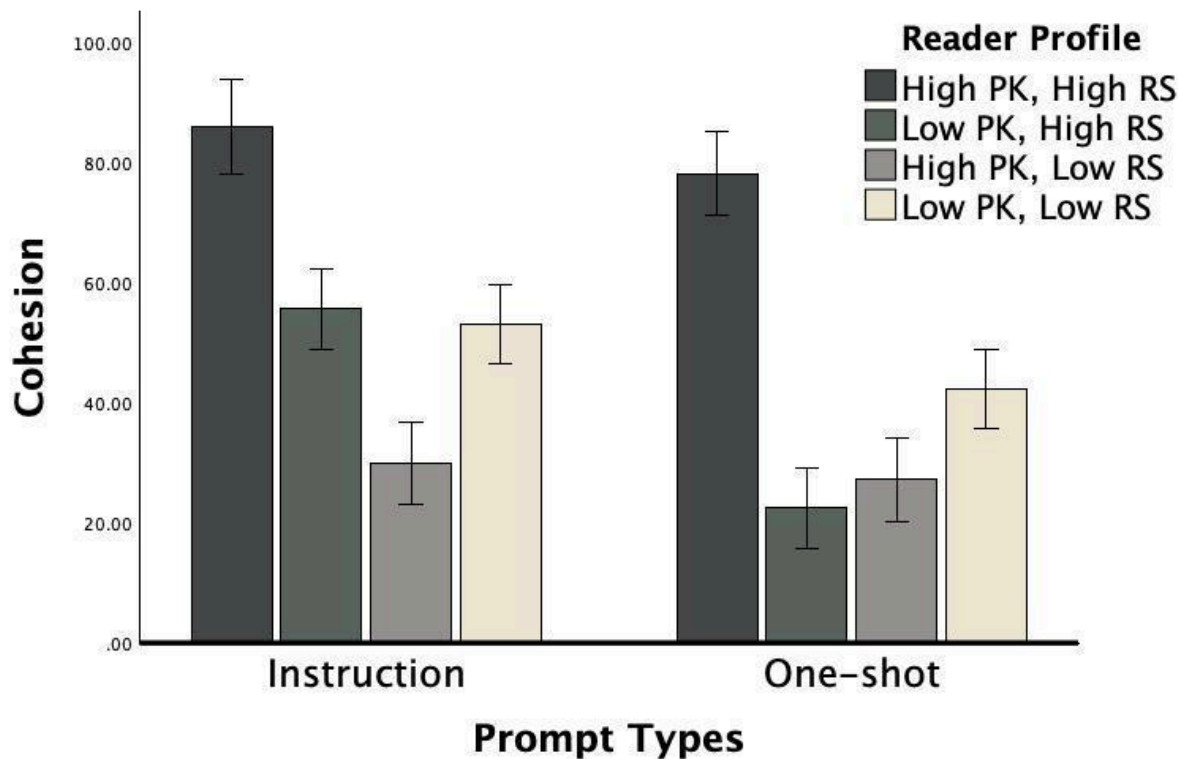
A 4 (Reader Profile: High RS/High PK, High RS/Low PK, Low RS/High PK, Low RS/Low PK) × 4 (LLM: ChatGPT-4, Gemini Pro 1.5, Claude 3.5 Sonnet, Llama 3.1) MANCOVA, controlling for word count, examined how linguistic features of LLM-modified history texts varied across reader profiles and prompting types.

Consistent with prior findings (Huynh & McNamara, 2025a; 2025b), LLMs systematically adjusted surface-level linguistic features to match reader ability. Texts adapted for skilled, high-knowledge readers (Profile 1) featured the highest syntactic and lexical sophistication, using academic writing style. Adaptations for less skilled or low-knowledge readers (Profiles 2, 3, and 4) featured shorter sentences, simpler vocabulary, and more concrete wording, indicating that LLMs successfully tailored text complexity to suit the readers' needs. These adjustments confirm that LLMs can scale personalization.

In contrast to the theory-aligned patterns observed for surface-level features (e.g., shorter sentences, simpler vocabulary, and more concrete wording for less skilled and low-knowledge readers), cohesion patterns did not align with comprehension theory (see Figure 1). Prior research suggests that low-knowledge readers benefit from cohesive texts, whereas high-

knowledge readers can learn more effectively from less cohesive texts (Ozuru et al., 2009). Contrary to this expectation, texts generated for lower-knowledge readers were less cohesive than those tailored for higher-knowledge readers. This reversal effect indicates that while LLMs can tailor surface features, they lack sensitivity to readers' discourse needs due to limited theoretical and contextual understanding. It is critical to provide additional scaffolding such as RAG-augmented technique to enhance LLM's adaptation.

Figure 1. Cohesion features as a function of reader profile and prompt types.



Discussion

LLMs systematically adjusted linguistic features to match reader characteristics, producing modifications that were generally appropriate for readers' needs. Texts for high-skill, high-knowledge readers exhibited greater syntactic and lexical complexity and a more academic writing style, whereas adaptations for low-knowledge or less skilled readers simplified sentence structures and vocabulary to enhance readability. These patterns confirm that LLMs can tailor surface-level complexity effectively and that NLP reliably quantifies these differences. However, cohesion patterns diverged from theoretical expectations. Although task-specific prompts improved linguistic sophistication, neither prompt type successfully aligned cohesion with comprehension theory. As a result, personalized texts seem to be more readable but they are not aligned with the readers' background knowledge in ways that are predicted to support comprehension. This mismatch highlights the need for knowledge-augmented prompting that integrates theoretical principles directly into prompt design. Embedding comprehension theory could enable adaptive systems to control both surface and discourse-level features more effectively (Huynh & McNamara, 2025a).

From a learning engineering perspective, this study represents a focused sub-cycle embedded within a broader adaptive personalization pipeline. The learning problem—ensuring that LLM-generated adaptations align with comprehension theory across reader profiles—motivated the design of theory-informed prompts. Implementation occurred through LLM-based text

generation, while NLP-based linguistic indices provided fine-grained, scalable measurements of alignment at both surface and discourse levels. Critically, the observed cohesion misalignment constitutes actionable design evidence rather than a terminal evaluation outcome. Within a learning engineering cycle, such evidence directly informs redesign decisions, including the integration of theory-augmented prompting strategies (e.g., RAG-based or self-reflective prompts) in subsequent iterations. This positioning distinguishes the framework from standalone validation or analytics approaches by embedding measurement within an ongoing improvement loop. As such, NLP functions not merely as an assessment tool but as an engineering control mechanism that enables rapid, data-driven refinement of AI behavior. This nested-cycle approach illustrates how learning engineering can support the development of adaptive, theory-aligned personalization systems that are both scalable and responsive to learner diversity (Baker et al., 2022; Goodell & Kolodner, 2023).

Extending prior studies, this work demonstrates the framework's generalizability to the history domain and identifies prompt design as a key factor in theory-aligned personalization. This study advances learning engineering by demonstrating how NLP serves as a scalable evaluation layer within adaptive learning systems. The framework enables real-time assessment of personalization quality and supports continuous, data-driven refinement across domains.

Although NLP can potentially provide a scalable method for validating LLM personalization, several open questions remain. How can these metrics be dynamically linked to learner performance data to determine whether linguistic alignment truly improves comprehension? What kinds of prompting architectures (e.g., RAG-based, self-reflective, or multi-agent systems) best translate theoretical constructs into generative behavior? Finally, how might these validation frameworks be generalized across languages and educational domains? Addressing these questions will advance the scientific foundations needed for reliable, explainable, and equitable AI-driven personalization.

Applying NLP validation in real educational systems also requires a robust infrastructure that connects text analytics with learner modeling and content generation, embedding NLP analytics directly into the design and delivery cycle. A scalable pipeline would (1) generate personalized content through LLMs, (2) apply NLP analytics in real time to evaluate linguistic alignment with learner profiles, and (3) feed these metrics back into the prompt or content generation process. Achieving this vision requires developing interoperable tools that connect LLM APIs, learner modeling systems, and NLP analytics dashboards capable of interpreting theory-based linguistic indices. Such integration will enable a transition from static, offline analyses to dynamic, real-time validation pipelines that ensure theory-aligned personalization at scale. Integrating these components transforms NLP validation from a research method into an operational mechanism for continuous system improvement—an essential step for learning engineering in practice.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

- Abbes, F., Bennani, S., & Maalel, A. (2024). Generative AI and gamification for personalized learning: Literature review and future challenges. *SN Computer Science*, 5(8), 1–12.
- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

- Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(1), 1–20.
- Dong, L., Tang, X., & Wang, X. (2025). Examining the effect of artificial intelligence in relation to students' academic achievement in classroom: A meta-analysis. *Computers and Education: Artificial Intelligence*, 8, 100400.
- Frantz, R. S., Starr, L. E., & Bailey, A. L. (2015). Syntactic complexity as an aspect of text complexity. *Educational Researcher*, 44(7), 387–393.
- Goodell, J., & Kolodner, J. (Eds.). (2023). *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond*. Taylor & Francis.
- He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., & Hasan, S. (2024). Does prompt formatting have any impact on LLM performance? arXiv preprint. arXiv:2411.10541. <https://doi.org/10.48550/arXiv.2411.10541>
- Huynh, L., & McNamara, D. S. (2025a). GenAI-Powered Text Personalization: Natural Language Processing Validation of Adaptation Capabilities. *Applied Sciences*, 15(12), 6791.
- Huynh, L., & McNamara, D. S. (2025b). Natural Language Processing as a Scalable Method for Evaluating Educational Text Personalization by LLMs. *Applied Sciences*, 15(22), 12128.
- Leong, J., Pataranutaporn, P., Danry, V., Perteneder, F., Mao, Y., & Maes, P. (2024). Putting things into context: Generative AI-enabled context personalization for vocabulary learning improves learning motivation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–15).
- Magliano, J. P., & Millis, K. K. (2004). Assessing reading skill with a think-aloud procedure. *Cognition and Instruction*, 21, 251–283.
- Martínez, P., Ramos, A., & Moreno, L. (2024). Exploring large language models to generate easy-to-read content. *Frontiers in Computer Science*, 6, 1394705.
- McNamara, D. S. (2021). Chasing theory with technology: A quest to understand understanding. *Discourse Processes*, 58(5–6), 422–448.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
- McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, 36(2), 222–233.
- Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, 47(1), 91–108.
- O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, 43(2), 121–152.
- Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19(3), 228–242.
- Pesovski, I., Santos, R., Henriques, R., & Trajkovik, V. (2024). Generative AI for customizable learning experiences. *Sustainability*, 16(7), 3034.

Potter, A., Shortt, M., Goldshtein, M., & Roscoe, R. D. (in press). Assessing academic language in tenth-grade essays using natural language processing. *Assessing Writing*.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., & Wang, G. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Ye, Q., Axmed, M., Pryzant, R., & Khani, F. (2023). Prompt engineering a prompt engineer. *arXiv preprint arXiv:2311.05661*.

Comments and Action Items for Authors



Open Repository for AI Models as Learning Engineering Components

Jim Goodell

AI

Design Patterns

Instrumentation

learning engineering

Machine Learning

Modeling

Modular Open Systems Architecture

Standards

HuggingFace.co is one of the most influential platforms for developers building, sharing and using machine learning models used by thousands of organizations, however, finding models applicable to learning is nearly impossible and the interface is highly technical. LearningFace.ai is offered as an alternative that making AI models more accessible to learning engineering teams. As the “hugging face for learning analytics” it offers a bridge between research and practice. It empowers researchers to get their work published and give learning engineers access to those discoveries for data-driven decision-making to optimize learning. The beta open access site is piloting a new approach to democratizing AI learner state detection as “learning engineering components”, packaged in a form that can be implemented in various learning systems. This poster will give a demonstration of learningface.ai and its use in learning engineering as a resource for learning engineering components.

Jim Goodell¹

¹ INFERable, a Public Benefit Corporation and IEEE Learning Technology Standards Committee; jim@inferable.app

Abstract.

Keywords: AI, Learning Engineering, Instrumentation. Modular Open Systems Architecture, Design Patterns, Standards, Machine Learning, Models.

Introduction

We developed LearningFace.ai an alternative to general purpose sites hosting model cards like huggingface. The site is designed as a public good resource to make learning analytics specific AI models more accessible to learning engineering teams. It is intended to offer a bridge between research and practice, to empower researchers to get their work published and give learning engineers access to those discoveries for data-driven decision-making to optimize learning. The beta open access site is piloting a new approach to democratizing AI learner state detection as “learning engineering components”, packaged in a form that can be implemented in various learning systems. This poster shown in figure 1 shows learningface.ai in its current state with placeholder learning cards.

Figure. 1.

Poster showing learningface.ai for hosting model cards for learning analytics and learner state detection models.

Open Repository for AI Models as Learning Engineering Components

Jim Goodell

INFERable, a Public Benefit Corporation; jim@inferable.app

HuggingFace.co is one of the most influential platforms for developers building, sharing and using machine learning models used by thousands of organizations, however, finding models applicable to learning is nearly impossible and the interface is highly technical. LearningFace.ai is offered as an alternative that making AI models more accessible to learning engineering teams. As the “hugging face for learning analytics” it offers a bridge between research and practice. It empowers researchers to get their work published and give learning engineers access to those discoveries for data-driven decision-making to optimize learning. The alpha open access site is piloting a new approach to democratizing AI learner state detection as “learning engineering components”, packaged in a form that can be implemented in various learning systems. The site currently has placeholder “model cards”. More research is need to test the use and usability of the approach.


Learning Face™ (alpha)
democratizing the power of learning analytics
a service of **INFERable™**

Logout

☰

Learning Analytics Models

Knowledge Tracing with Deep Learning
Knowledge Tracing
Employs deep neural networks to model student knowledge progression. -- This advanced Knowledge Tracing model utilize...
Info Use

Automated Feedback Generation for Essays
Feedback & Assessment
Provides automated feedback on student essays based on linguistic features. -- This model analyzes student-written essays,...
Info Use

Student Engagement Prediction from Video Logs
Engagement Analytics
Predicts student engagement levels by analyzing their interactions with learning videos. -- This model uses features extracte...
Info Use

Early Warning System for At-Risk Students
Student Success
Predicts students who are at risk of academic failure based on various indicators. -- This predictive model integrates data from variou...
Info Use

Sentiment Analysis of Student Forum Posts
Social Learning Analytics
Analyzes the sentiment expressed in student posts within online discussion forums. -- This Natural Language Processing (NLP) model...
Info Use

Adaptive Learning Path Recommendation Engine
Personalized Learning
Recommends personalized learning paths based on student performance and goals. -- This recommendation engine uses student...
Info Use

Detection of Off-Task Behavior in Online Learning
Engagement Analytics
Identifies students exhibiting off-task behavior during online learning sessions. -- This model analyzes student interactions (e.g., mouse...
Info Use

Concept Map Generation from Learning Materials
Knowledge Representation
Automatically generates concept maps from textual learning resources.
Info Use

Peer Assessment Quality Prediction
Feedback & Assessment
Predicts the quality and helpfulness of peer assessments. -- This model analyzes the text and features of peer reviews to predict thei...
Info Use

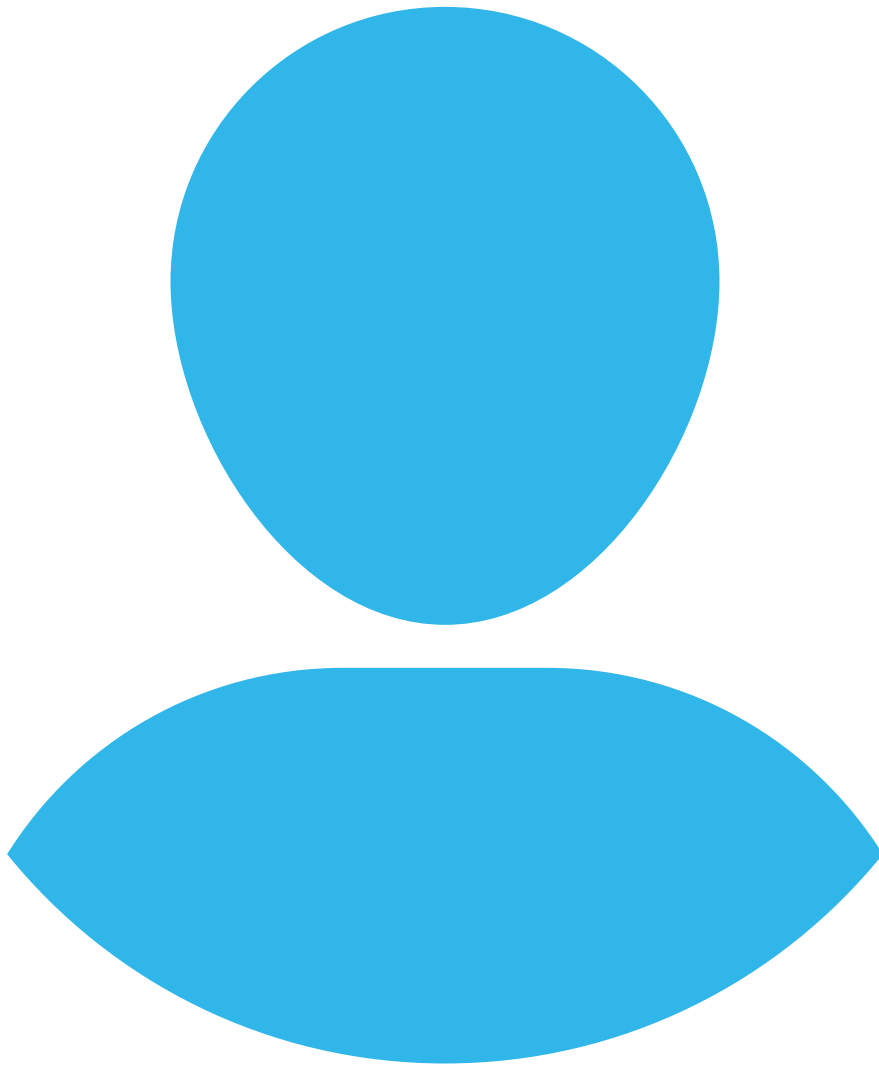
LearningFace.ai

Discussion and Conclusion

More research is needed on the generalizability of models made available as learning engineering components as well as more translation of research to formats that could be used by practitioners via learningface.ai. We invite the research community to work with us to bridge the research-practice gap for learning analytics and learner state detection methods and models by publishing learning cards to the site and using the site as a resource.

References

- Goodell, J., & Kolodner, J. (Eds.). (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Routledge.
- IEEE. (n.d.). IEEE P2247.2 Draft Standard for Adaptive Instructional Systems Interoperability.
- IEEE. (n.d.). IEEE P2997 Draft Standard for Enterprise Learner Record (ELR).
- IEEE. (2023). IEEE 1484.20.3-2022: IEEE Standard for Learning Technology: Data Model for Sharable Competency Definitions (SCD).
- IEEE. (2024). IEEE 1484.2-2024: IEEE Recommended Practice for Learning and Employment Record (LER) Ecosystems.
- IEEE. (2025). IEEE 2881-2025: IEEE Standard for Learning Metadata Terms (LMT).
- IEEE. (2023). IEEE 9274.1.1-2023: Standard for Experience API (xAPI).
- INFERable.app. (n.d.). INFERable.app [Website].
- learningface.ai. (n.d.). learningface.ai [Website].



Jim Goodell



Optimizing Language-Focused Writing Feedback from Large Language Models through Prompt Engineering

Large Language Models

prompt engineering

Writing Feedback

Yu Tian, Andrew Potter, Jessica Early, Steve Graham, Parul Patil, Katerina Christhlf, Motahareh Darvishpour Ahandani, and Danielle McNamara

Arizona State University; {ytian126, ahpotter, jessica.early, steve.graham, ppatil57, kchristh, mdarvis2, dsmcnamara}@asu.edu

High-quality, individualized feedback is essential for supporting the development of students' academic writing skills, yet instructors face persistent challenges delivering such feedback at scale in writing-intensive courses. Large language models (LLMs) have the potential to provide timely and personalized feedback, but the pedagogical quality of this feedback is highly sensitive to the design of the prompt (Dai et al., 2024). Despite growing adoption of AI writing support tools, little empirical research has examined which prompting strategies most effectively elicit accurate and actionable feedback.

This project investigates how different prompt engineering strategies affect the quality and instructional effectiveness of language-focused writing feedback. Four prompting conditions were implemented and evaluated:

1. Baseline Prompt: providing a generic request for writing feedback.
2. Rubric-Augmented Prompt: embedding an analytic rubric targeting language features.
3. NLP-Feature-Informed Prompt: incorporating natural language processing (NLP) metrics such as lexical diversity and syntactic complexity.
4. Combined Prompt: integrating rubric criteria with NLP metrics.

Undergraduate students (N = 113) completed a timed narrative writing task and a revision session informed by LLM feedback assigned under one prompting condition. A panel of writing instructors independently evaluated the generated feedback and students' drafts and revisions using custom-developed rubrics.

Initial analyses suggest that augmented prompts (i.e., rubric-based, NLP-informed, and combined prompts) generate more development-oriented feedback focused on improving language sophistication and clarity, while the baseline prompt tends to produce predominantly accuracy-focused comments. We expect these augmented prompts to lead to more meaningful, development-related revisions and greater improvements in students' language quality.

This study advances learning engineering by providing empirical guidance on how prompt design shapes the pedagogical value of LLM-generated writing feedback. The findings of this study can inform the development of scalable writing support systems that adapt to learners' needs and prompt deeper engagement with language features that matter for academic writing. More broadly, the work demonstrates a systematic, evidence-driven approach in learning engineering (Baker et al., 2022) to optimizing human-AI interaction for complex learning tasks such as academic writing.

References

Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*.

Dai, W., Tsai, Y. S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7, 100299.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.



Quality Assessment Through Learning Engineering: An Evaluation Rubric of LLM-Generated Multiple-Choice Questions

Katerina Christhlf, Linh Huynh, Yu Tian, Shubham Chakraborty, Micah Watanabe, & Danielle S. McNamara

Automated Question Generation

Multiple Choice Questions

Psychometrics

Accurately assessing reading comprehension is essential for monitoring student progress, yet creating high-quality multiple-choice questions is time-consuming for instructors. AI-generated questions offer a promising solution but raise concerns about question quality. To address this gap, we developed a rubric for human raters to evaluate the quality of automatically generated comprehension questions. The rubric incorporates established criteria for question design and documented weaknesses of AI-generated items, including awkward, wordy phrasing and conspicuous correct answers. Through iterative testing with two human raters scoring 200 AI-generated multiple-choice questions, we refined scoring categories and scales, achieving weighted kappa reliability of .79 or higher. The final rubric includes seven categories assessing answer correctness, distractor incorrectness, topic relevance, writing clarity, distractor linguistic features, distractor semantic plausibility, and distractor semantic uniqueness. This tool supports future research comparing question-generation systems and exploring automated question-quality evaluation.

Katerina Christhlf, Linh Huynh, Yu Tian, Shubham Chakraborty, Micah Watanabe, and Danielle S. McNamara

Arizona State University; {kchristh, lthuyh1, ytian126, schak103, mwatana5, dsmcnama}@asu.edu

Abstract.

Keywords: Automated Question Generation, Multiple Choice Questions, Psychometrics.

Challenge

Accurate assessment of reading comprehension is essential for instructors to evaluate students' progress and adjust instruction. Multiple-choice questions (MCQs) are widely used in education since they offer an efficient and objective, and method for large-scale assessments (Haladyna et al., 2002); however, they can be difficult and time-consuming for instructors to create. High-quality MCQs contain unambiguously correct answers, plausible but incorrect distractors, clear and concise wording, alignment with learning objectives, and emphasis on key information rather than minor details (Haladyna et al., 2002). As such, a key educational challenge is how best to support instructors in rapidly creating relevant, appropriately challenging MCQs for learner practice and assessment.

Recent advances in Generative Artificial Intelligence (GenAI) have enabled automatic generation of MCQs at scale, reducing the burden on instructors. Yet, there are concerns regarding the quality of these questions. Automatically generated questions tend to have certain characteristics (Gorgun & Bulut, 2024; Zhou & Li, 2025). Although AI-generated questions are excellent in spelling and grammar, they often overemphasize trivial details, include awkward or overly verbose wording, and produce distractors that are shorter or less developed than the correct answer. These limitations highlight the need for a human-rating rubric that systematically evaluates the quality of AI-generated questions. There have been previous human evaluations of MCQs, but they often focused on surface details or were overly broad (e.g., Cheung et al., 2023; Olney, 2023). The objective of this work is to introduce a domain general question-evaluation rubric designed to assess the quality of MCQs.

Rubric Design

Our team created an initial rubric that was iteratively refined through three rounds of practice scoring sessions. The finalized rubric includes seven categories (see Appendix A): Answer Correctness, Distractor Incorrectness, Topic Relevance, Writing Clarity, Distractor Linguistic Features, Distractor Semantic Plausibility, and Distractor Semantic Uniqueness. Answer Correctness indicates that the “correct answer is clearly correct to those who have comprehended the text”, while Distractor Incorrectness indicates that the “distractors are clearly incorrect to those who have comprehended the text.” Because GenAI-generated questions rarely include inaccurate answers or accurate distractors, answer correctness and distractor incorrectness have binary rating options: 0 or 1. Other categories capture known challenges for LLM-generated questions, including overemphasis on trivial details, awkward or overly verbose language, and implausible or insufficiently distinct distractors. Therefore, the remaining categories were scored on a 4-point scale to allow nuanced quality judgments while avoiding the midpoint bias common in 3-point scales. Topic Relevance indicates that the question “addresses a central concept in the text”. Writing Clarity indicates that the question “stem and response options are concise and unambiguous”. While other rubrics often separate clarity into several categories, such as grammar, spelling, and fluency, we condensed clarity into a single category, because artificially generated questions rarely include errors in grammar or spelling. The features of the distractors were split into three categories: linguistic features, semantic plausibility, and semantic uniqueness, to reflect three common yet separate issues with artificially generated questions. Distractor Linguistic Features refers to how similar the language of the distractors is to the correct response, including length, syntactic structure, and style. Distractor Semantic Plausibility refers to the extent to which the meanings of the distractors are plausible, while Distractor Semantic Uniqueness refers to the extent to which each distractor conveys a unique meaning.

Preliminary Outcomes

Two subject-matter experts used the rubric to evaluate a set of 200 AI-generated MCQs. Results of post-discussion weighted kappas reached .79 or higher for every category (see Appendix B), indicating the rubric supported raters in establishing good reliability when scoring these key features of MCQs.

Conclusion

This project demonstrates an application of learning engineering by systematically designing and refining a key assessment artifact, a human-rating rubric for MCQ quality, within a larger effort to evaluate AI-generated questions. Grounded in learning engineering’s emphasis on evidence-based design, the rubric was developed through initial research on key question quality considerations, followed by three cycles of use, feedback, and revision; design decisions were informed by rater agreement and scoring discussions. This work aligns with the creation phase of a broader learning engineering cycle by enabling consistent data collection about known strengths and weaknesses of GenAI-produced MCQs. It will be used to support subsequent phases of our larger project evaluating questions generated by different LLMs and multi-agent systems. At the same time, the rubric development itself can be viewed as a smaller, nested learning engineering cycle involving goal definition, artifact design, practical testing, and evidence-based refinement. Looking ahead, we are also exploring the potential

to automate this rubric using LLMs, natural language processing, or machine learning methods, enabling the carefully developed scoring categories from this rubric to be used at a larger scale.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., Co, M. T. H. (2023). ChatGPT versus human in generating medical graduate exam questions – An international prospective study. medRxiv. <https://doi.org/10.1101/2023.05.13.23289943>

Gorgun, G., & Bulut, O. (2024). Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. Education and Information Technologies, 29(18), 24111–24142. <https://link.springer.com/article/10.1007/s10639-024-12771-3>

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Olney, A. M. (2023). Generating multiple choice questions from a textbook: LLMs match human performance on most metrics. Proceedings of the Workshop on Empowering Education with LLMs. <https://eric.ed.gov/?id=ED630037>

Zhou, H., & Li, L. (2025). Qadg: Generating question–answer–distractors pairs for real examination. Neural Computing & Applications, 37, 1157–1170. <https://doi.org/10.1007/s00521-024-10658-5>

Appendix A

Rubric to Evaluate Multiple-Choice Question Quality

		0- Disagree		1- Agree	
Answer Correctness					
Correct answer is clearly correct to those who have comprehended					
Distractor Incorrectness					
Distractors are clearly incorrect to those who have comprehended the text.					
		1- Disagree	2- Neutral	3- Agree	4- Strongly Agree

Overall Quality The question and answer choices are clearly worded, concise, and unambiguous. Addresses key concepts central to the text.	Topic Relevance Addresses a central concept in the text	Question contains trivial or irrelevant information	Question contains a key sub-point, but no main ideas	Question contains a key sub-point that connects to main ideas	Entire question focused on a main idea or key concept
	Writing Clarity Stem and options concise & unambiguous	Excessively wordy or unclear	Wordiness or other slightly impacts comprehension, mild ambiguity, or awkward syntax	A few redundancies or minor wordiness	Fully clear and concise
Distractor Plausibility Distractors are plausible to those who have not comprehended the text	Linguistic Features All are similar in linguistic features (e.g., length, number of phrases)	Distractors are visibly different from correct answer (e.g., length by 12+ characters).	Distractors all have small differences, OR at least one distractor has major differences from correct answer.	Distractors have relatively same length and features to correct answer	All distractors are uniform in linguistic features
	Semantic Plausibility All are semantically plausible (e.g., similar topic, no absolute terms, information is not more well known than the correct answer)	All distractors mildly implausible OR two distractors very implausible	Two distractors mildly implausible OR one distractor very implausible	One distractor is mildly plausible	All distractors plausible
	Semantic Uniqueness Each distractor is unique in meaning	All distractors very similar	Multiple distractors slightly similar or two very similar	Two distractors are wrong in similar way	All distractors unique

Appendix B

Post-Discussion Weighted Kappas Per Category

Scoring Category	Weighted Kappa
Answer Correctness	1
Distractor Incorrectness	0.79
Topic Relevance	0.97

Writing Clarity	0.96
Distractor Linguistic Features	0.84
Distractor Semantic Plausibility	0.83
Distractor Semantic Uniqueness	0.89



Reasoning LLMs are Competent Courseware Reviewers

Ryan P. Dwyer & David J. Yaron

Automated evaluation

Content validation

LLM

Expanded Abstract

Introduction and Approach

There is robust causal evidence that learning by doing is more effective than reading or watching video alone (Koedinger et al., 2018), so evidence-based online courseware centers student problem-solving with feedback. The REAL CHEM general chemistry courseware integrates over 1,000 questions into the text compared to just 176 in the traditional Open Stax Chemistry 2e textbook. This evidence-based design presents a challenge for course designers, however, as it is difficult for subject matter experts and student reviewers to identify and correct errors in this large quantity of targeted feedback. Scalable methods are needed to help authors review and improve large sets of activities efficiently.

Reasoning LLMs enable these questions and their feedback to be checked for errors accurately and inexpensively. We developed a system to convert each question to markdown and pass it to a reasoning LLM for review. To verify and improve the quality of the LLM's output, we built an automated evaluation system using an LLM-as-a-judge architecture. This iterative design-test-refine process represents a nested learning engineering cycle within the broader REAL CHEM development effort.

Findings and Implications

For the automated evaluation system, we selected and identified errors (or verified no errors were present) in 60 representative REAL CHEM questions. The performance of several gpt-5 family models was measured against this set, with gpt-5-mini with medium reasoning effort and a carefully targeted prompt template achieving 88 percent accuracy. The results show that model size, reasoning depth, and prompt design all matter: limited reasoning reduces sensitivity to real errors, while targeted prompts significantly outperform generic ones. The automated evaluation system also enables iterative prompt improvement.

This approach offers a practical, low-cost method for improving large-scale instructional content. Modifications to this general approach have been used to assign alt text and learning objectives where those were missing. In REAL CHEM, the system has already improved hundreds of questions in the existing courseware.

References

Koedinger, K. R., Scheines, R., & Schaldenbrand, P. (2018). Is the Doer Effect Robust Across Multiple Data Sets? International Educational Data Mining Society.



ReQUESTA: A Hybrid Agentic Framework for Generating Cognitively Diverse Multiple-Choice Questions

Yu Tian, Shubham Chakraborty, Linh Huynh, Katerina Christhlf, Micah Watanabe, & Danielle S. McNamara

Cognitive Variety

Hybrid Agentic Framework

Large Language Models

Question Generation

This study presents ReQUESTA, a hybrid agentic framework that integrates LLM-powered and rule-based agents to generate multiple choice questions (MCQs) with distinct cognitive focuses: text-based, inferential, and main-idea. To provide an initial validation of the framework, expert raters evaluated 100 ReQUESTA-generated MCQs using a cognitive classification rubric to assess alignment between intended and perceived cognitive categories. The results indicate a high level of agreement between system-assigned and expert-assigned labels (agreement rate = 0.95), suggesting that ReQUESTA can reliably instantiate targeted cognitive distinctions in question generation. These findings offer preliminary evidence of the framework's capacity to support cognitively diverse and pedagogically meaningful assessment design. ReQUESTA's modular and hybrid design supports scalable, iterative, and evidence-based assessment development. Future work will include psychometric validation and expansion to additional cognitive categories such as application-level questions.

Introduction

Multiple-choice questions (MCQs) remain a widely used format for automated assessment in education due to their efficiency, objective scoring, and scalability across large student cohorts (Simkin & Kuechler, 2005). However, developing high-quality MCQs requires substantial time and expertise from instructors and assessment designers. Consequently, there is a growing need for automated MCQ generation systems that can reduce instructor workload and maintain the pedagogical integrity of assessments while supporting scalable and evidence-based learning evaluation.

Recent developments in transformer-based large language models (LLM) suggest that automatically generating high-quality MCQs has become increasingly feasible. For example, Olney (2023) found that LLM-generated questions drawn from textbook content achieved parity with human-authored items on most metrics of item quality. Likewise, Cheung et al. (2023) showed that ChatGPT generated MCQs that were broadly comparable to those drafted by faculty, exhibiting similar clarity, specificity, discriminative power, and suitability for graduate-level examination.

Nonetheless, key challenges persist. Specifically, alignment of generated items to higher cognitive levels (e.g., inferential or analytical reasoning) remains difficult. Recent evaluations (Scaria et al., 2024), drawing on Bloom's taxonomy (2020), indicate that LLMs perform well for lower-level (recall) items. However, performance and alignment degrade for higher-order items unless supported by additional control mechanisms, such as advanced prompting, verification classifiers, or specialized refinement agents.

Accordingly, we introduce ReQUESTA, a hybrid multi-agent framework that integrates rule-based components with LLM-powered agents to generate cognitively diverse MCQs across three levels: (i) text-based (recall), (ii) inferential, and (iii) main-idea (synthesis). The system decomposes the resource-intensive MCQ authoring process into a sequence of specialized subtasks, each handled by a dedicated agent (e.g., Planner, Question Generators, Evaluator). By combining the generative strengths of LLMs with the precision and control of rule-based logic, the system establishes a scalable, consistent, and pedagogically grounded workflow for automated MCQ generation in educational assessment contexts.

Background and Framework

Agentic systems have emerged as a practical architecture to leverage LLMs on complex reasoning tasks that involve multiple steps (e.g., Christie et al., 2025; Lu et al., 2025). In these systems, agents (e.g., fine-tuned models, external software tools, custom python scripts) may be specialized for distinct responsibilities (e.g., planning, generation, verification, and orchestration), and the frameworks coordinate these roles to decompose difficult tasks into tractable subtasks and to combine deterministic rule-based processing with stochastic LLM outputs.

Recent developments demonstrate the growing success of multi-agent pipelines for automated question generation. Pawar et al. (2024) introduced a LangChain-based multi-step framework using Gemini LLMs to generate and evaluate quizzes from structured instructional text, producing instructor-usable questions with minimal human revision. Similarly, Mucciaccia et al. (2025) presented an LLM-based agentic system combining prompt engineering and automated evaluation, which generated high-quality MCQs while substantially reducing faculty workload. Despite these advances, existing frameworks are not yet able to reliably control the cognitive complexity of generated questions, and alignment with higher-order cognitive levels (e.g., inferential or synthesis questions) continues to be an open challenge.

ReQUESTA Workflow Design

ReQUESTA's design (illustrated in Appendix A) extends beyond conventional automated question generators that produce primarily recall-based items. It strategically analyzes the source text to identify key concepts, inferences, and overarching ideas, enabling the generation of text-based, inferential, and main-idea MCQs based on user specifications. The system is composed of multiple interacting agents, described below.

Core Architecture

1. Preprocessor: A rule-based agent that uses spaCy's sentence segmentation to divide the input text into coherent chunks, preserving contextual integrity while ensuring manageable input size for downstream processing.
2. Planner: An LLM-powered agent prompted with structured, stepwise instructions to (1) summarize each text segment, (2) extract key ideas and inferences, and (3) construct a high-level question generation plan according to user inputs (e.g., desired number of each question type).
3. Controller: A rule-based coordination agent implemented with custom Python scripts. It interprets the Planner's output and assigns generation tasks along with the relevant text segments to appropriate Question Generator agents based on the specified configuration.
4. Question Generators (Text-based, Inferential, Main-idea). Three LLM-powered agents guided by detailed prompts and examples to produce MCQs aligned with the cognitive focus defined in the plan: key concepts for text-based questions, logical reasoning for inferential questions, and overarching understanding for main-idea questions.
5. Evaluator: An LLM-powered agent that applies a pedagogically grounded evaluation checklist (e.g., item clarity, stem-answer alignment, distractor plausibility) to assess the quality of each generated question. Questions meeting the quality criteria are approved for output, while those requiring improvement are automatically returned to the relevant generator for revision.
6. Formatter: A rule-based agent implemented through a set of Python scripts that finalizes the question output by shuffling answer options, standardizing formatting, and applying consistent labeling across all generated items.

Option-Shortening Module

The Option-Shortening Module was developed to ensure uniformity in the length and structure of MCQ options. It detects options that are noticeably longer than others and adjusts them to maintain stylistic consistency without altering meaning. The module consists of four collaborating agents:

1. Syntax Analyzer. An LLM-powered agent that identifies common syntactic patterns across the four options. These patterns guide the generation of revised options to ensure linguistic consistency and structural uniformity.
2. Length Determiner. A rule-based agent implemented with custom Python scripts that establishes acceptable length thresholds and flags options exceeding these limits.
3. Candidate Generator. An LLM-powered agent that produces concise candidate revisions for any overlong option while retaining its original meaning.
4. Candidate Selector. An LLM-powered agent guided by rule-based metrics (e.g., word count, cosine similarity) to evaluate the candidates and select the version that best meets length and stylistic criteria.

Evaluation

The purpose of this evaluation was to provide an initial validation of ReQUESTA's ability to generate MCQs that instantiate intended cognitive focuses (text-based, inferential, and main-idea questions). Specifically, the evaluation examined the degree of alignment between system-assigned cognitive labels and expert judgments, thereby assessing the construct validity of the framework's cognitive differentiation.

Materials and Question Generation

An open-access, college-level textbook, *Introduction to Anthropology* (OpenStax, 2024), was used as the source material for question generation. Ten chapters were randomly selected from the textbook. The chapters averaged 1,835 words in length ($SD = 361$), ranging from 1,309 to 2,480 words.

Each chapter was processed using ReQUESTA with GPT-4o as the major LLM engine. For each chapter, ReQUESTA generated 10 questions, yielding a total of 100 questions. The system was configured to produce four text-based, four inferential, and two main-idea questions per chapter, consistent with the targeted distribution of cognitive focuses.

Expert Evaluation

All 100 generated MCQs were randomized, and their system-assigned cognitive labels were removed before evaluation to prevent potential bias. Two subject-matter experts independently classified each question using the rubric described in Appendix B. Following independent coding, inter-rater agreement was assessed. The two raters achieved a high level of agreement, with an agreement rate of 94% and a Cohen's kappa of 0.91, indicating almost perfect agreement according to the benchmark proposed by Landis and Koch (1977). The raters then met to discuss and resolve disagreements, and the reconciled classifications were used for subsequent analysis.

To examine the extent to which ReQUESTA-generated questions aligned with their intended cognitive focuses, the system-assigned labels were compared with the expert-assigned consensus labels. Standard classification metrics—accuracy, precision, recall, and F1-score—were computed to summarize overall alignment and category-specific patterns. In this context, accuracy reflects the overall proportion of questions whose intended cognitive focus matched expert judgments, while precision and recall characterize alignment within individual cognitive categories. The F1-score, defined as the harmonic mean of precision and recall, provides a balanced summary of category-level alignment.

Results

Example MCQs of different cognitive focuses generated by ReQUESTA are provided in Appendix C. Table 1 reports the alignment between ReQUESTA’s intended cognitive labels and expert-assigned classifications, and Figure 1 presents the corresponding confusion matrix.

Across all categories, ReQUESTA achieved an overall agreement rate of 0.95, indicating that expert raters classified 95% of the generated questions into the same cognitive categories intended by the system. This high level of alignment provides initial evidence that the framework can reliably operationalize distinct cognitive focuses in MCQ generation.

Category-level analyses showed consistently strong alignment across text-based, inferential, and main-idea questions. In particular, main-idea questions exhibited perfect alignment, with all items classified by experts as intended. Minor discrepancies were observed primarily between text-based and inferential categories. As shown in the confusion matrix, four out of 40 questions intended to be text-based were classified by experts as inferential. Qualitative inspection suggests that these items, while grounded in textual content, required additional reasoning or synthesis, thereby blurring the boundary between

Figure. 1. *A confusion matrix comparing ReQUESTA-generated and human-assigned labels.*

recall-oriented and inferential processing.

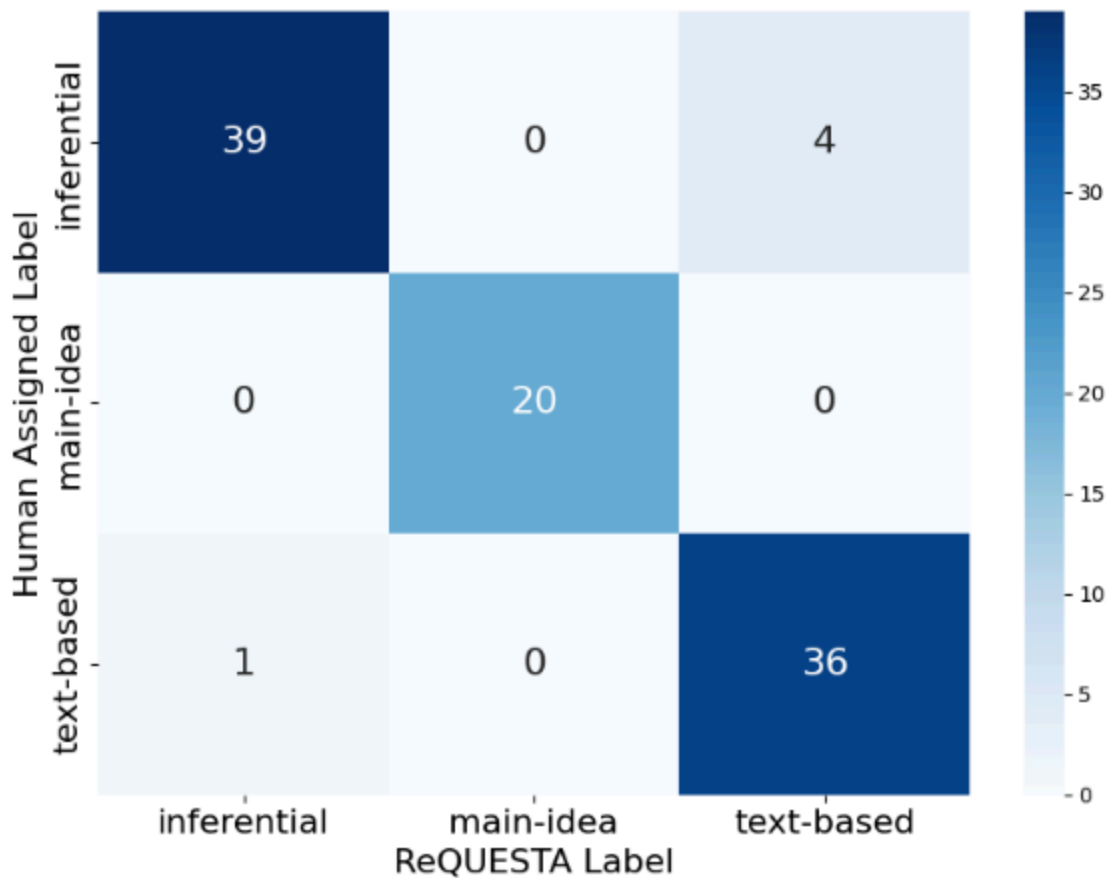


Table 1. Performance Metrics for ReQUESTA Question Type Label against Human Assigned Labels

Label	Precision	Recall	F1 Score
text-based	0.90	0.973	0.935
inferential	0.975	0.907	0.940
main-idea	1.00	1.00	1.00
Overall	0.95	0.95	0.95

Discussion

The evaluation provides initial evidence that ReQUESTA can reliably instantiate distinct cognitive focuses (text-based, inferential, and main-idea) in generated multiple-choice questions, as reflected in the high level of alignment between system-assigned labels and expert judgments. These findings support the construct validity of the framework's cognitive differentiation mechanism. In particular, the complete alignment observed for main-idea questions suggests that ReQUESTA effectively operationalizes global comprehension and synthesis processes that are central to higher-order cognitive assessment. The limited discrepancies between text-based and inferential questions are informative rather than problematic, highlighting the inherently graded boundary between surface-level recall and reasoning-based processing. For example, some questions labeled by the system as text-based required examinees to evaluate the relative relevance or implications of explicitly stated information, prompting expert raters to classify them as inferential. Such cases underscore the continuum nature of cognitive demands in assessment tasks and point to opportunities for further refinement of cognitive category definitions and agent prompting strategies.

The study also demonstrates the effectiveness of using agentic architectures to tackle cognitive complex tasks such as MCQ generation. Specifically, MCQ authoring involves heterogeneous subtasks: content retrieval and segmentation, identification of key concepts, stem composition, distractor generation, stylistic normalization, and pedagogical validation. By assigning these subtasks to dedicated agents, ReQUESTA demonstrates the capacity to move beyond fact-recall items toward more cognitively demanding questions that assess inference and synthesis. This represents an advancement over conventional question generators, which often emphasize surface linguistic quality rather than cognitive depth (Ch & Saha, 2018).

Furthermore, the hybrid and modular nature of ReQUESTA supports scalable, iterative, and evidence-driven processes for designing and refining learning systems, aligning with contemporary characterizations of learning engineering (Baker et al., 2022). The framework facilitates continuous cycles of design, implementation, evaluation, and refinement through the integration of human expert judgment and state-of-the-art intelligent agents. In this sense, ReQUESTA can be conceptualized as a nested learning engineering cycle embedded within a broader assessment design workflow, with the goal of reducing instructor workload while enhancing the cognitive rigor and pedagogical alignment of automated assessments.

Conclusion and Future Directions

This study introduced ReQUESTA, a hybrid agentic framework that integrates rule-based and LLM-powered agents to generate multiple-choice questions with distinct cognitive focuses. Through an initial expert-alignment evaluation, the study provides evidence that ReQUESTA can reliably instantiate intended cognitive categories (text-based, inferential, and main-idea), highlighting the framework's capacity to operationalize cognitively differentiated assessment design. These findings

underscore the potential of hybrid, agentic AI systems to support scalable and pedagogically meaningful assessment generation.

Future work will focus on systematic evaluation of question quality through human judgments and psychometric analyses to assess the validity, reliability, and instructional value of the generated items. Continued development will also aim to extend ReQUESTA's capabilities to produce a broader range of cognitive categories, including application and evaluation questions, thereby enhancing its educational utility and alignment with established learning taxonomies.

Acknowledgments

The research reported here was supported by Arizona State University (ASU), ASU Learning Engineering Institute and the Institute of Education Sciences, U.S. Department of Education, through Grants R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent views of ASU, the Institute of Education Sciences, or the U.S. Department of Education.

References

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*. <https://doi.org/10.1037/tmb0000058>
- Bloom, B. S. (2010). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Ch, D. R., & Saha, S. K. (2018). Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1), 14-25.
- Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., ... & Co, M. T. H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong SAR, Singapore, Ireland, and the United Kingdom). *PloS one*, 18(8), e0290691.
- Christie, S. T., Rafferty, A. N., Lee, Z., Cutler, E., Tian, Y., & Almoubayyed, H. (2025, July). An agentic framework for real-time pedagogical plot generation. In *International Conference on Artificial Intelligence in Education* (pp. 309-316). Cham: Springer Nature Switzerland.
- Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.
- Lu, P., Chen, B., Liu, S., Thapa, R., Boen, J., & Zou, J. (2025). Octotools: An agentic framework with extensible tools for complex reasoning. *arXiv preprint arXiv:2502.11271*.
- Mucciaccia, S. S., Paixão, T. M., Mutz, F. W., Badue, C. S., de Souza, A. F., & Oliveira-Santos, T. (2025, January). Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 2246-2260).
- Olney, A. M. (2023). *Generating Multiple Choice Questions from a Textbook: LLMs Match Human Performance on Most Metrics*. Grantee Submission.

OpenStax. (2024). Introduction to anthropology. OpenStax, Rice University. <https://openstax.org/details/books/introduction-anthropology>.

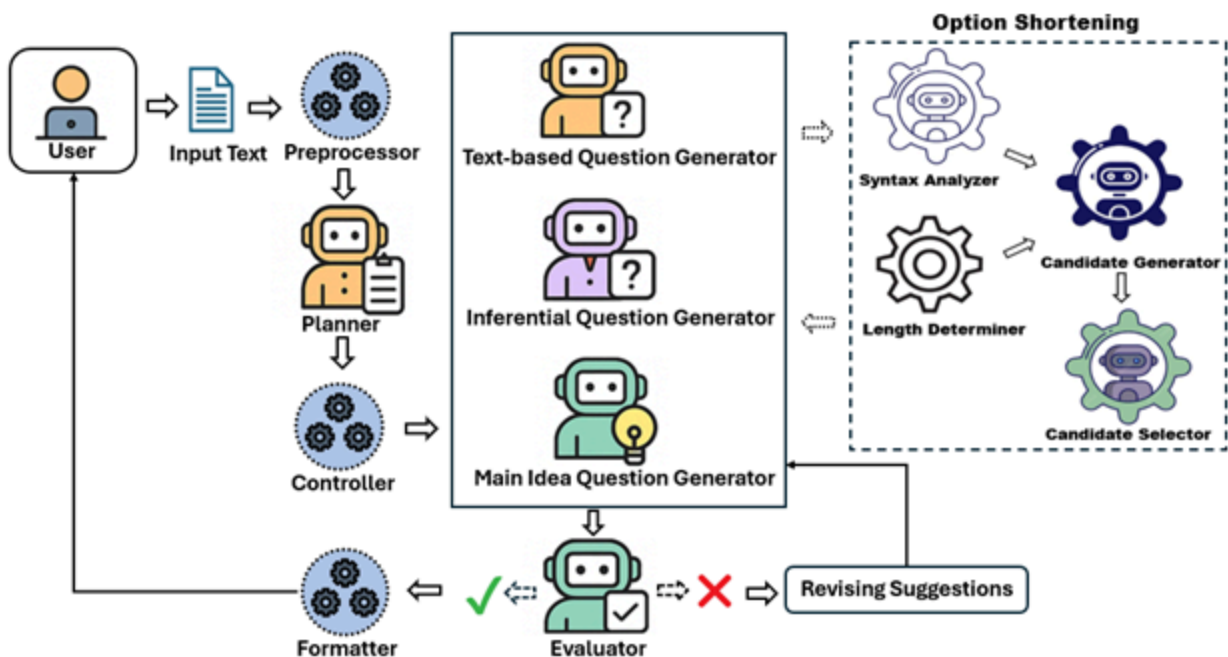
Pawar, P., Dube, R., Joshi, A., Gulhane, Z., & Patil, R. (2024, August). Automated Generation and Evaluation of MultipleChoice Quizzes using Langchain and Gemini LLM. In 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT) (Vol. 1, pp. 1-7). IEEE.

Scaria, N., Dharani Chenna, S., & Subramani, D. (2024, July). Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In International Conference on Artificial Intelligence in Education (pp. 165-179). Cham: Springer Nature Switzerland.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection?. Decision Sciences Journal of Innovative Education, 3(1), 73-98.

Appendix A

Figure. A1. Overview of the ReQUESTA workflow. The process begins with a user-uploaded file (.pdf, .docx, or .txt). The Preprocessor extracts and segments the text, which is then analyzed by the Planner to identify key concepts and inferences and to construct a question generation plan. This plan is parsed by the Controller, which coordinates and assigns tasks to the three Question Generators responsible for producing text-based, inferential, and main-idea questions concurrently. The Option-Shortening Module standardizes answer options by shortening noticeably longer alternatives to ensure consistent length and structure. The Evaluator assesses each question's quality and determines whether it requires revision by the relevant generator or is ready for output. Finally, the Formatter organizes and delivers the approved questions to the user interface.



Appendix B

Table B1.

Rubrics for Assessing Different Question Types and Their Cognitive Focuses.

Can the exact answer to this question be located from one sentence in the text?

Yes- Text-based

No- Inferential or Main idea

Text-based	Inferential	Main idea
Information is directly stated from the text	Information is not explicitly stated, must be inferred	Overall point, purpose, or central topic of the text
Answer can be found verbatim or paraphrased in one sentence	Answer choice is a combination of ideas from more than one sentences in the text	
	What can be inferred... How would the author feel about...	What is the main idea... Which statement best summarizes...

Appendix C

Example MCQs Generated by ReQUESTA

Text-based

What role do values and beliefs play in shaping society?

- A) They enforce legal standards and judicial processes within a culture.
- B) They suggest what is considered good and bad, beautiful and ugly.
- C) They regulate individuals' emotions and personal feelings.
- D) They determine the distribution of resources among societal groups.

Inferential

How does the Sapir-Whorf hypothesis explain the influence of language differences on cultural perception and behavior?

- A) Language structures can lead to different interpretations of gender roles across cultures.
- B) The adoption of new vocabulary is uniform across cultures, reflecting a shared global culture.
- C) Words have attached meanings that universally influence cognition, regardless of cultural context.

D) Language innovations in technology influence all cultures similarly, despite prior linguistic differences.

Main-idea

What is the main idea of the passage?

A) Anthropology has remained unchanged since the 19th century and still relies on “armchair anthropology”.

B) Anthropology's evolution includes diverse perspectives and the move from indirect to direct research methods.

C) Feminist anthropology was the sole contributor to modern anthropological practices.

D) Modern anthropology focuses only on ethnographic methods without any contributions from other disciplines.



Social and Emotional Dimensions of Generative AI Use

Rebekah Jongewaard, Lindsey McCaleb, Nicole Oster, Emmanuel Adeloju, & Punya Mishra

Artificial Intelligence

Education

Generative AI

Technology

Extended Abstract

Reports of generative AI's (GenAI's) alarming influence on users' social, psychological, and affective states proliferate. GenAI attunes to specific contexts and individual users' emotional states, desires, and vulnerabilities (Nash, 2024), yet much extant research on GenAI in education treats it as a relationally neutral tool. As learning engineers, our aim is to design human-centered solutions to the challenges posed to education in an AI-saturated society.

To better understand these challenges, we explored the following questions in one micro-cycle nested within the "Challenge" phase of our larger project (Craig et al., 2025):

1. How do university students use GenAI?
2. What range of affective experiences do students report in their interactions with GenAI, and how do these experiences impact their academic engagement, social interactions, and personal well-being?

Preliminary results from thematic analysis of an initial student interview include (a) anxiety related to uncertain or ignored academic expectations for AI use and (b) creative uses of AI for personal purposes. We present an agenda for continuing research on GenAI that centers the social and emotional well-being of students. By better understanding these increasingly urgent issues we can work toward uses of GenAI that prioritize human flourishing.

Acknowledgments

Funding for this research was provided in part by MLFTC: Division of Educational Leadership and Innovation at ASU.

References

Craig, S. D., Avancha, K., Malhotra, P., Gorman, J. C., Verma, V., Likamwa, R., ... & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale.

Nash, B. L. (2024). Love and Learning in the Age of Algorithms: How Intimate Relationships with Artificial Intelligence May Shape Epistemology, Sociality, and Linguistic Justice. *Reading Research Quarterly*, 59(4), 624–631. Academic Search Ultimate. <https://doi.org/10.1002/rrq.549>



Socio-Emotional Learning in AI K-12 Guidance and Policy Documents: A Gap Analysis

Emmanuel Adeloju, Lindsey McCaleb, Rebekah Jongewaard, Nicole Oster, & Punya Mishra

Artificial Intelligence

K-12 Education

Socioemotion

As generative AI becomes increasingly integrated into K–12 classrooms, it poses distinct socio-emotional risks that existing policies inadequately address. This study analyzed 36 institutional, state, and international AI guidance documents using the CASEL framework and five inductively generated SEL-informed categories: Anthropomorphization & Mental Model, Emotional Attachment & Parasocial Relationships, Engineered Manipulation & Supernormal Stimuli, Social Isolation & Relationship Displacement, and Developmental Vulnerability. Content analysis revealed uneven attention across categories: Developmental Vulnerability and Engineered Manipulation dominated, appearing in 88% of documents, while Emotional Attachment received minimal coverage (8%). The documents highlighted risks including diminished student agency, algorithmic bias, erosion of peer and teacher relationships, and the potential for parasocial engagement with AI. These findings reflect the need for policies that scaffold critical evaluation, maintain human oversight, promote relational and social-emotional development, and mitigate exploitation of developmental vulnerabilities.

Introduction

As conversational AI advances, interactions with it increasingly blur the human-machine boundary. In a cross-sectional study comparing AI chatbot responses to physician responses to patient health questions, chatbot responses were rated as more empathetic and of higher quality than those from physicians (Ayers et al., 2023). Moreover, people also often attribute human-like qualities and emotions to AI (Shevlin, 2024; Mishra & Oster, 2025), a tendency that can shape emotional responses and decision-making, especially among adolescents, whose cognitive and socioemotional capacities are still developing (Dumontheil, 2024; Larson et al., 2002). These emerging capabilities of GenAI raise concerns about how AI-mediated interactions might influence students' emotional regulation, relationship skills, and social development. Yet, while most K12 schools now embrace the use of AI, there is limited attention paid to the socioemotional risks that accompany students' engagement with GenAI. Hence, educational policies must address AI's relational and psychological dimensions, as well as the implications of its usage. Furthermore, policies must emphasize "Supporting Human Processes," which centers human-in-the-loop decision-making (Baker et al., 2022). This study examines how K-12 education policies and guidance documents address the socioemotional implications of student-AI interaction. Drawing on the Collaborative for Academic, Social, and Emotional Learning (CASEL, 2020) framework, which identifies five competencies: self-awareness, self-management, social awareness, relationship skills, and responsible decision-making, we analyzed 36 institutional and organizational AI policy documents. Synthesizing across studies (Bakir & McStay, 2025; Epley et al., 2007; Mishra, 2024, 2025a, 2025b), we inductively generated five thematic areas for evaluating the policy and guidance documents which are: (1) Anthropomorphization & Mental Model,

where students attribute human qualities and consciousness to AI, (2) Emotional Attachment & Parasocial Relationships: One-sided relationships, emotional bonds, and treating AI as companions, (3) Engineered Manipulation & Supernormal Stimuli: deliberate design to exploit human psychology which are exaggerated or artificial versions of natural cues that trigger our evolved instincts more powerfully than the original things those instincts were designed for. (4) Social Isolation & Relationship Displacement: Erosion of social skills, reduced human connection, and (5) Developmental Vulnerability: automatic cognitive bias, social-emotional needs. Our research questions, hence, are: (1) How are education policies addressing the socio-emotional risks of GenAI in K-12 settings? (2) What socio-emotional elements should K-12 AI policies include to mitigate these risks?

Methodology

This study employed qualitative content analysis (Mayring, 2000) to examine how social and emotional learning (SEL) considerations are articulated in K-12 Artificial Intelligence (AI) guidance documents. Qualitative content analysis was selected to enable the structured interpretation of the document texts while remaining sensitive to both explicit statements and implicit meanings related to SEL.

Search Strategy

Document identification began with the AI for Education State AI Guidance webpage, which yielded 34 state-level guidance documents. To broaden coverage, particularly for international and organizational perspectives, additional documents were identified through LinkedIn posts, news articles, and direct searches of state and organizational websites, resulting in 2 country-level and 11 organization-level documents. In total, 47 records were identified. One duplicate document was removed prior to screening, leaving 46 documents for review. These records were screened using predefined inclusion and exclusion criteria, resulting in the exclusion of 10 documents. The final analytic sample consisted of 36 guidance documents, all of which were included in the qualitative content analysis. Throughout the process, documents were cataloged in a spreadsheet that evolved from a reference repository into an analytic tool, incorporating fields for metaphors, skill emphases beyond foundational principles, and the degree to which social-emotional learning (SEL) components were addressed.

Inclusion and Exclusion Criteria

Documents were selected based on predefined inclusion and exclusion criteria. To be included in the analysis, documents were required to (a) be published in English, (b) focus on the use of artificial intelligence in K–12 educational contexts, and (c) originate from reputable governmental, educational, or organizational sources. Documents were excluded if they did not contain any explicit or implicit references to social and emotional learning (SEL) components.

Findings

The analytic approach used in this study involved five main SEL categories, each subdivided into conceptually distinct subcategories. Anthropomorphization & Mental Model captures misunderstandings of AI's nonhuman nature, operationalized through Human-like Attribution, where students treat AI as possessing emotions, intentions, or moral agency. Emotional Attachment and Parasocial Relationships reflects affective misalignment, represented by One-sided Attachment, in which learners form emotional bonds with AI despite its lack of reciprocity. Engineered Manipulation and Supernormal Stimuli addresses psychologically exploitative design and includes Persuasive Feature Response, referring to AI features that shape behavior, blur privacy boundaries, or influence thinking; Reduced Self-Regulation, indicating difficulties managing attention or impulses in AI-mediated environments; and Targeted Vulnerability Exploitation, where AI leverages developmental limits such as novelty-seeking or difficulty distinguishing fact from fiction. Social Isolation and Relationship

Displacement captures erosion of human social engagement through Reduced Peer Interaction, marked by diminished peer-to-peer engagement; Simplified Social Practice, reflecting flattened or conformist social interaction; and Support Network Erosion, describing weakened help-seeking relationships with teachers, caregivers, or mentors. Finally, Developmental Vulnerability encompasses age-related susceptibilities, including Concurrent SEL Strain, where multiple SEL competencies are challenged simultaneously; Skill-Gap Exploitation and Overestimation of Capability, involving misplaced trust in AI accuracy, neutrality, or judgment; and Diminished Agency, characterized by deference to AI and reduced student initiative or decision-making. These subcategories were inductively generated during the coding process through iterative analysis of the documents

Content analysis of the 36 policy and guidance documents revealed substantial variation in how SEL-related concerns about generative AI were addressed. Two main categories dominated the corpus: Developmental Vulnerability and Engineered Manipulation & Supernormal Stimuli, each appeared in 88% (n=32) of documents. In contrast, Emotional Attachment & Parasocial Relationships was notably underrepresented, appearing in only 8% of documents (n=3). Social Isolation & Relationship Displacement appeared in 41% of documents (n=15), while Anthropomorphization & Mental Model was present in 36% (n=13).

Within-category patterns revealed concentrated attention on specific concerns. For Developmental Vulnerability, Diminished Agency was most prevalent (n=23, 63%), followed by Skill-Gap Exploitation & Overestimation of Capability (47%, n=17), while Concurrent SEL Strain appeared in only 13% of documents (n=5). The Engineered Manipulation & Supernormal Stimuli category showed even greater concentration: Targeted Vulnerability Exploitation dominated at 83% (n=30), while Reduced Self-Regulation appeared in only 2% of documents (n=1), representing the lowest frequency of any subcategory. Social Isolation & Relationship Displacement subcategories showed more balanced but modest coverage: Support Network Erosion (22%, n=8), Simplified Social Practice (n=6, 16%), and Reduced Peer Interaction (n=5, 13%).

Anthropomorphization & Mental Model

Across 13 documents addressing Anthropomorphization & Mental Model, they consistently foregrounded the need to delimit AI's non-human nature, anticipate student misconceptions, and adopt preventive communication practices. The documents explicitly stated that AI lacks core human qualities, emphasizing that it has no "consciousness, emotions, and self-awareness" (Delaware Department of Education) and is "not artificial brains or sentient life forms with human characteristics like free will, self-awareness, and emotions" (Oregon Department of Education), operating instead "without comprehension, awareness, or intent" (OECD, 2025) and as a system "extremely good at predicting the logical sequence of words" but "still not 'sentient'" (Connecticut Commission for Educational Technology). At the same time, the documents acknowledged risks of anthropomorphic misconceptions, including evidence that "students believe AI may have a real person inside it" (Rhode Island Department of Elementary and Secondary Education) and concerns that AI's ability to "mimic human patterns of communication" (Arizona Institute for Education and the Economy) and simulate "human companionship" (Connecticut Commission for Educational Technology) may blur relational boundaries. In response, several sources emphasized intentional communication as a safeguard, stressing that students must be "clearly informed that it is a chatbot and not a human" (North Carolina Department of Public Instruction) and that AI systems should signal that "its social interaction is simulated and that it has no capacities of feeling or empathy" (European Union, 2022), reinforcing accurate mental models that avoid attributing human feelings or understanding to AI.

Developmental Vulnerability

The 32 documents addressing Developmental Vulnerability, guidance concentrated on three analytically generated subcategories, Diminished Agency, Concurrent SEL Strain, and Skill-Gap Exploitation & Overestimation of Capability, with uneven depth. Diminished Agency was most prominent, with the documents repeatedly warning that "overreliance on AI models can lead to undercutting the learning process and abandoning human discretion and oversight" (Delaware Department of Education) and that "excessive dependence on AI tools may hinder the development of critical thinking, problem-solving,

and independent learning skills” (Rhode Island Department of Elementary and Secondary Education), prompting strong emphasis on maintaining human control, opt-out mechanisms, and learners as “critical consumers” of AI (Utah State Department of Education). Concurrent SEL Strain appeared more sparingly, but framed AI risks as compounding across competencies, linking metacognitive awareness of AI’s influence on “behaviors, thoughts, and learning processes” (OECD) with broader concerns about well-being, safety, and harms such as AI-enabled bullying (Delaware Department of Education). Skill-Gap Exploitation & Overestimation of Capability highlighted developmental risks of treating AI as authoritative, cautioning that it is “NOT: A replacement for educators... [or] a perfect or self-correcting system” (Massachusetts Department of Elementary and Secondary Education) and that outputs may be “confident sounding but factually incorrect” (Missouri Department of Elementary and Secondary Education), while stressing that AI “cannot comprehend nuance or contextual information unless explicitly told in the prompt” (Missouri Department of Elementary and Secondary Education) and that younger students are “not cognitively ready to completely understand AI or critically evaluate content” (North Carolina Department of Public Instruction). Together, these findings depict developmental vulnerability as centered on protecting agency, recognizing overlapping SEL pressures, and countering inflated assumptions about AI capability through human oversight and critical evaluation.

Emotional Attachment & Parasocial Relationships

Addressed in only three documents, Emotional Attachment & Parasocial Relationships received limited but pointed attention, centering on the relational risks posed by AI systems that simulate human interaction without the capacity for reciprocity. These documents cautioned against the rise of AI companion-like use, noting that “the trajectory of digital ‘companions’ is steep, with people of all ages already engaged in chatting with AI-powered bots that simulate behaviors of human companionship” and warning that such interactions “may further isolate young people and negatively impact emotional well-being” (Connecticut Commission for Educational Technology). At the same time, the guidance emphasized AI’s fundamental relational limits, stressing that “AI is limited in the ability to read emotional cues, understand cultural contexts fully, or establish genuine human connections with students” (Colorado Education Initiative) and that it “cannot truly empathize with human emotions” (North Dakota Department of Public Instruction). Together, these sources frame parasocial attachment not as a widespread phenomenon but as a foreseeable risk when simulated companionship obscures AI’s inability to engage in authentic emotional exchange.

Engineered Manipulation & Supernormal Stimuli

Engineered Manipulation & Supernormal Stimuli was widely addressed, with the 32 documents noting AI’s subtle influence through persuasive design, hidden algorithms, and personalized recommendations. Sources emphasized that AI “works in ways that are usually not visible or easily understood by users” (European Union, 2022) and that it may “collect your personal information without you knowing it” (Washington Office of Superintendent of Public Instruction). Persuasive Feature Response can steer attention and reinforce biases, with learners encouraged to “identify bias” in AI outputs (Wisconsin Department of Public Instruction) and recognize that AI “will inherently reflect societal biases in the data” (Utah State Board of Education; West Virginia Department of Education). While Reduced Self-Regulation was only indirectly referenced, Targeted Vulnerability Exploitation was extensively discussed, highlighting risks such as difficulty distinguishing fact from fiction, amplification of inequities, and exploitation of specific student vulnerabilities. The documents cautioned that AI “can result in new forms of inequalities or discrimination and exacerbate existing ones” (European Union, 2022) and that “these tools can easily behaviorally profile people such that individuals can be tracked through their online behavior without ever revealing their identity” (Rhode Island Department of Elementary and Secondary Education). Together, these sources frame AI as a powerful but often invisible actor, requiring critical reflection, oversight, and inclusive design to mitigate harm.

Social Isolation & Relationship Displacement

Social Isolation & Relationship Displacement was widely noted as a risk of AI use, with 15 documents highlighting how technology can reduce authentic human connection, narrow social perspectives, and erode critical support networks. Reduced

Peer Interaction was linked to diminished interpersonal skill development, with sources emphasizing that AI “cannot stand in for peer to peer exchange that is key to the development of interpersonal competencies” (Colorado Education Initiative)) and that its use “may further isolate young people and negatively impact emotional well-being” (Connecticut Commission for Educational Technology). Simplified Social Practice was similarly cautioned against, as AI may foster “an artificial conformity of perspectives and creating impediments to understanding the real world” (InnovateOhio) and “hinder constructive dialogue and understanding between students” (California Dept of Education). Support Network Erosion was also highlighted, noting that AI can weaken teacher-student relationships and depersonalize learning, with warnings that “over-reliance on AI-generated content may lead to a false sense of understanding...less of a connection between the student and the teacher” (Rhode Island Department of Elementary and Secondary Education) and that “AI can result in new forms of inequalities or discrimination and exacerbate existing ones” (European Union, 2022). Collectively, these sources frame AI as a force that can disrupt relational development, social-emotional well-being, and equitable educational experiences unless carefully monitored and balanced with authentic human interaction.

Discussion

The findings reveal that SEL-related guidance on generative AI emphasizes both the cognitive and relational dimensions of student risk, with attention unevenly distributed across categories. Developmental Vulnerability and Engineered Manipulation & Supernormal Stimuli dominated the corpus, reflecting widespread concern about learners’ diminished agency, susceptibility to biased outputs, and exploitation of developmental limitations. Furthermore, the documents consistently cautioned that overreliance on AI “may hinder the development of critical thinking, problem-solving, and independent learning skills” (Rhode Island Department of Elementary and Secondary Education) and that AI will reflect societal biases in the data (Utah State Board of Education; West Virginia Department of Education). While Emotional Attachment & Parasocial Relationships received limited attention, sources highlighted that AI interactions “may further isolate young people and negatively impact emotional well-being” (Connecticut Commission for Educational Technology) and emphasized that AI “cannot truly empathize with human emotions” (North Dakota Department of Public Instruction). Together, these findings suggest that policymakers recognize AI’s potential to influence learning, thinking, and social-emotional development, but coverage is uneven and concentrated on certain high-profile risks.

Social Isolation & Relationship Displacement further highlight the relational consequences of AI adoption, showing how technology can erode peer engagement, flatten social practice, and weaken critical support networks. The documents emphasized that AI “cannot stand in for peer to peer exchange that is key to the development of interpersonal competencies” (Colorado Education Initiative) and warned that “over-reliance on AI-generated content may lead to a false sense of understanding...less of a connection between the student and the teacher” (Rhode Island Department of Elementary and Secondary Education). Furthermore, risks of conformity and narrowed perspectives were also mentioned, cautioning that AI may foster “an artificial conformity of perspectives and creating impediments to understanding the real world” (InnovateOhio) and “hinder constructive dialogue and understanding between students” (California Dept of Education). These patterns indicate that, alongside cognitive and developmental concerns, policymakers are increasingly attentive to how AI can disrupt relational and equity-oriented aspects of education, emphasizing the need for oversight, human-centered engagement, and inclusive design to safeguard both social-emotional well-being and equitable learning experiences.

Future Directions

To mitigate socio-emotional risks of generative AI in K-12 settings, future policy and guidance documents should explicitly incorporate strategies aligned with the five SEL-informed categories identified in this study. For Anthropomorphization & Mental Model, policies should ensure students clearly understand that AI “does not possess consciousness, emotions, or intent” and that its social interaction is simulated. For Emotional Attachment & Parasocial Relationships, guidance should

promote structured, human-mediated engagement, such as peer collaboration, teacher-facilitated discussions, and reflection exercises, to prevent students from forming one-sided emotional bonds with AI.

Regarding Engineered Manipulation & Supernormal Stimuli, policies should embed transparency and bias-awareness measures, requiring AI outputs to include prompts or disclaimers that highlight potential algorithmic influence, and supporting metacognitive activities that help learners critically evaluate AI-generated content and recommendations. To address Social Isolation & Relationship Displacement, guidance should maintain human connection through teacher check-ins, family involvement, and opportunities for collaborative learning, reducing over-reliance on AI for social engagement. Finally, for Developmental Vulnerability, policies should scaffold student agency, support decision-making, and emphasize age-appropriate critical evaluation, ensuring that learners can recognize AI limitations, avoid overestimation of capability, and manage concurrent social-emotional demands effectively.

Collectively, embedding these socio-emotional elements in K-12 AI policies offers actionable pathways to safeguard students' learning, relational development, and well-being while enabling responsible and informed engagement with AI systems.

Reference

- Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., ... & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6), 589-596.
- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*. <https://doi.org/10.1037/tmb0000058>
- Bakir, V. & McStay, A. (2025). Move fast and break people? Ethics, companion apps, and the case of Character.ai. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5159928
- Collaborative for Academic, Social, and Emotional Learning - CASEL (2020). What Is the CASEL Framework? A framework creates a foundation for applying evidence-based SEL strategies to your community.
- Dumontheil, I. (2014). Development of abstract thinking during childhood and adolescence: The role of rostrolateral prefrontal cortex. *Developmental cognitive neuroscience*, 10, 57-76.
- Epley, N., Waytz, A., & Cacioppo, J.T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864-886
- <https://casel.org/fundamentals-of-sel/what-is-the-casel-framework>
- <https://punyamishra.com/2025/05/02/engineered-for-attachment-the-hidden-psychology-of-ai-companions/>
- European Commission: Directorate-General for Education, Youth, Sport and Culture, Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators, Publications Office of the European Union. (2022). <https://data.europa.eu/doi/10.2766/153756>
- Larson, R. W., Moneta, G., Richards, M. H., & Wilson, S. (2002). Continuity, stability, and change in daily emotional experience across adolescence. *Child development*, 73(4), 1151-1165.
- Mayring, P. (2021). Qualitative content analysis: A step-by-step guide.

- Mishra, P. (2024, December). The ELIZA Effect-ion. Punya Mishra. <https://punyamishra.com/2024/12/01/the-eliza-effect-ion/>
- Mishra, P. (2025a, February). The Attribution Problem: Why we can't stop seeing ourselves in AI. Punya Mishra. <https://punyamishra.com/2025/02/11/the-attribution-problem-why-we-cant-stop-seeing-ourselves-in-ai/>
- Mishra, P., & Oster, N. (2025, March). Human or AI? Understanding the Learning Implications of Anthropomorphized Generative AI. In Society for Information Technology & Teacher Education International Conference (pp. 502-507). Association for the Advancement of Computing in Education (AACE).
- Mishra, P. (2025b, May). Engineered for Attachment: The Hidden Psychology of AI Companions. Punya Mishra.
- OECD (2025). Empowering learners for the age of AI: An AI literacy framework for primary and secondary education (Review draft). OECD. Paris.
- Shevlin, H. (2024). All too human? Identifying and mitigating ethical risks of Social AI.

Appendix

All US state documents analyzed can be found [here](#)

European Commission: Directorate-General for Education, Youth, Sport and Culture, Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators, Publications Office of the European Union, 2022, <https://data.europa.eu/doi/10.2766/153756>



The Difficult Conversations Bot: Findings on Fostering Empathy and Reflective Communication Among Faculty and Staff

Catheryn Reardon & James Dunnigan

AI-Mediated Role-Play

Human-Centered AI

learning engineering

This paper presents final findings from the design and evaluation of the Difficult Conversations Bot, an AI-mediated role-play system developed to support empathetic communication among higher education faculty and staff. Grounded in Learning Engineering and informed by humanistic, social cognitive, and adult learning theories, the system was evaluated using a pre-post mixed-methods design. Quantitative survey results demonstrated consistent positive shifts in confidence, intentional perspective-taking, psychological safety, reflective practice, and perceptions of AI as a learning partner. Qualitative findings highlighted the value of practicing a difficult conversation in a psychologically safe rehearsal simulation with structured reflection. Together, these results suggest that AI-mediated role-play, when designed through a nested learning engineering cycle, can function as an adaptive, human-centered learning system for complex interpersonal skill development.

Keywords: Learning Engineering, AI-Mediated Role-Play, Human-Centered AI.

Introduction

Faculty, staff, and K–12 teachers routinely navigate emotionally complex interpersonal conversations with students, colleagues, and, in K–12 contexts, parents. These conversations often include responding to student distress, managing peer conflict, addressing academic integrity concerns, and communicating expectations within layered organizational structures. These interactions are not peripheral to higher education work; they are central to student persistence, faculty effectiveness, staff well-being, and institutional climate. Yet professional development opportunities rarely provide structured, low-stakes environments in which educators can rehearse difficult conversations and reflect on their communicative choices before encountering them in real contexts.

The absence of rehearsal opportunities is especially consequential given the increasing complexity of higher education roles. Faculty and staff are expected to manage emotionally charged conversations related to mental health, equity and inclusion, performance feedback, and institutional change, often under time pressure and with limited guidance. Traditional professional development formats, such as policy briefings or one-time workshops, tend to emphasize procedural compliance rather than the nuanced interpersonal judgment required in these difficult conversations.

To address this gap in training, the Difficult Conversations Bot was developed as an AI-mediated role-play system designed to support empathy, reflective communication, and interpersonal awareness while rehearsing a difficult conversation. The project is grounded in Learning Engineering, which conceptualizes educational technologies as adaptive learning systems developed through iterative, data-informed cycles (Baker et al., 2022), and Arizona State University's Principled Innovation framework, which emphasizes human-centered, ethically grounded design (Arizona State University, 2022). Rather than

positioning AI as a productivity tool, this work frames AI-mediated role-play as a learning system that supports deliberate practice, meaningful interpretation, and reflection in emotionally complex professional contexts.

Framework and Design Process

The Difficult Conversations Bot was designed using a nested Learning Engineering cycle, embedding local design decisions and evaluation activities within a broader research and implementation effort (Craig et al., 2025; Totino & Kessler, 2024). This nested structure enabled the research team to iteratively refine the system while maintaining traceability between theoretical commitments, design choices, and learner data. Within Learning Engineering, effectiveness emerges not from isolated features but from alignment across theory, design, implementation, and evaluation.

The system integrates multiple complementary theoretical perspectives. Rogers' (1959) humanistic theory shaped the conversational tone and emphasis on empathy, unconditional positive regard, and authentic listening. Bandura's (1986) social cognitive theory guided the emphasis on rehearsal, modeling, and reflective self-efficacy, positioning communication skill development as experiential and iterative rather than declarative. Guided by Mezirow's (1991) transformative learning theory, structured reflection prompts were incorporated to support critical reflection and perspective shifts following each simulated interaction. Appreciative Inquiry (Cooperrider & Srivastva, 1987) framed strengths-based feedback that encouraged learners to identify effective strategies rather than focus solely on deficits. Bot training and implementation drew from the Reasoned Action Approach (Fishbein & Ajzen, 2010) and organizational change models such as ADKAR (Hiatt, 2006), shaping role-play structure to help learners practice intentional perspective-taking, action planning, and readiness for change in simulated interactions.

Design followed an iterative design-build-test-refine process consistent with learning engineering practice. Early prototypes were reviewed internally to evaluate emotional realism, ethical alignment, and conversational coherence. Pilot testing with faculty and staff informed refinements to scenario framing, instructional guidance, pacing, and post-dialogue reflection prompts. Each iteration drew on learner feedback and survey data to support systematic refinement across nested learning-engineering cycles and transparent design decisions.

Implementation and Methods

Following Institutional Review Board (IRB) approval, the Difficult Conversations Bot was tested using a pre–post mixed-methods design aligned with learning engineering evaluation practices. Participants were faculty and staff affiliated with a large public university who voluntarily engaged with the Difficult Conversations Bot as part of a research study on AI-mediated professional learning.

Participants completed a pre-survey assessing baseline experience with generative AI, confidence in navigating difficult conversations, and expectations for AI-supported learning. This baseline data established participants' starting conditions and helped distinguish learning-related change from novelty effects or lack of prior exposure. Participants then engaged asynchronously with one or more AI-mediated role-play scenarios via a web-based interface.

Scenarios were designed to reflect common professional challenges in higher education, such as responding to a distressed student, addressing ongoing performance concerns with a colleague, or navigating supervisory expectations. The AI system generated contextually adaptive responses while maintaining emotional tone and conversational continuity. After completing the role-play, participants completed a post-survey consisting of 1–5 Likert-scale items and several open-ended reflection prompts.

Quantitative measures assessed confidence, intentional perspective-taking, psychological safety, reflective practice, and perceptions of AI as a learning partner. Survey constructs were grounded in the principles of Principled Innovation, emphasizing human-centered design, ethical reflection, psychological safety, and responsible AI use. Items were designed to capture learners' perceptions of reflective engagement, perspective shifting, and confidence in applying insights from the simulated interactions. Quantitative data were analyzed descriptively to examine pre–post shifts, while qualitative responses were analyzed using inductive thematic analysis to identify patterns relevant to learning processes and system design.

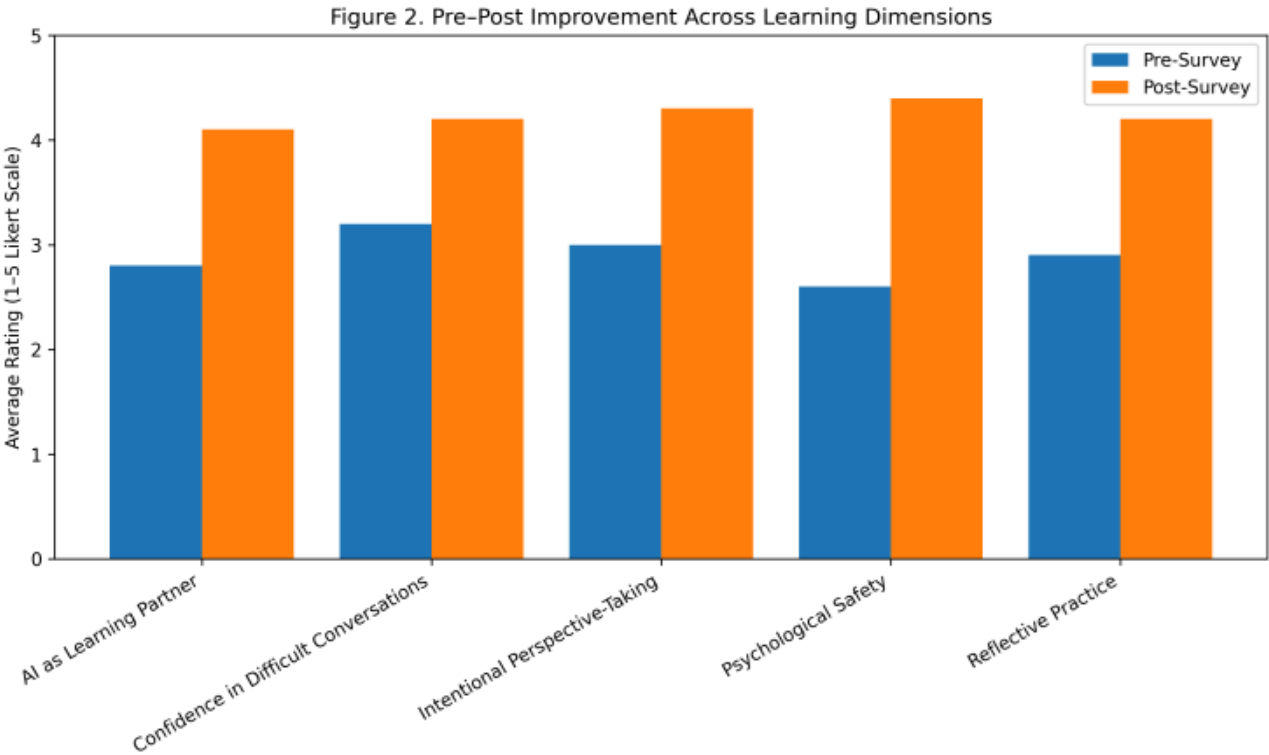
Findings and Insights

Pre-survey results indicated moderate to high baseline familiarity with generative AI and existing confidence in handling difficult conversations, alongside uncertainty regarding AI’s role in supporting relational and emotionally complex learning. A total of 20 participants completed the pre-survey, and many initially viewed AI primarily as an efficiency or information-support tool rather than as a medium for interpersonal skill development.

Sixteen participants completed the post-survey following engagement with the AI-mediated role-play (three participants did not complete the post-survey due to incomplete participation in the full activity sequence). Post-survey results demonstrated consistent positive shifts across all measured dimensions, including confidence, intentional perspective-taking, psychological safety, reflective practice, and perceptions of AI as a learning partner. These shifts suggest learning gains attributable to the design of the role-play as a reflective learning system rather than novelty effects. Participants reported greater intentionality in language choice and increased awareness of how conversational framing influenced outcomes.

Qualitative findings reinforced quantitative trends. Participants described the psychologically safe rehearsal environment as particularly valuable, noting that the ability to experiment without real-world consequences reduced anxiety and supported deeper reflection. Several participants highlighted moments of perspective shift in which they recognized assumptions or habits that might otherwise go unnoticed in real interactions.

Figure 1.
Pre–Post Improvement Across Learning Dimensions. This figure presents side-by-side comparisons of average pre-survey and post-survey ratings across five learning dimensions aligned with learning engineering principles. Ratings were collected using a 1–5 Likert scale (1 = strongly disagree, 5 = strongly agree). Pre-survey bars represent baseline participant perceptions prior to engagement with the AI-mediated role-play, while post-survey bars represent perceptions following the intervention. Across all dimensions, post-survey ratings were higher than pre-survey ratings, indicating consistent perceived improvement following engagement with the learning system.



Discussion and Implications

Interpreted through a learning engineering lens, these findings suggest that AI-mediated role-play can function as an adaptive, human-centered learning system for interpersonal skill development. The pre–post changes suggest that reflective simulation can meaningfully support learning even for experienced professionals, particularly when psychological safety, learner agency, and structured reflection are intentionally built into the experience.

The nested learning engineering approach enabled alignment between design intent, implementation, and evaluation. Pre–post survey data provided actionable evidence to guide refinement, while qualitative insights illuminated learner-system interaction patterns relevant to future iterations. Ethical considerations emphasized by Principled Innovation, including psychological safety, transparency, and learner agency, remain central, particularly when deploying AI in emotionally sensitive learning contexts.

From a practice perspective, AI-mediated role-play offers a scalable supplement to traditional professional development by providing low-risk opportunities for deliberate practice. Limitations include reliance on self-reported data and a single institutional context, underscoring the need for continued evaluation across settings.

Conclusion and Next Steps

This study demonstrates how learning-engineered AI-mediated role-play can augment professional learning by providing scalable, low-risk opportunities for reflective practice. Final pre–post findings suggest that the Difficult Conversations Bot supported increased confidence, intentionality, and reflective awareness in navigating difficult conversations.

Looking ahead, future work will replicate and extend this evaluation with larger and more diverse participant samples across institutional contexts to explore broader applicability and longer-term impact in workplace settings. In parallel, the research team intends to pursue external funding to advance this work toward an avatar-based, multimodal learning system that leverages natural conversational models to increase social presence, emotional expressiveness, and perceived authenticity of interactions. This direction directly addresses learning engineering grand challenges related to scaling complex skill development, designing human-centered AI systems that support judgment and empathy, and creating learning environments that approximate real-world complexity without sacrificing psychological safety. By integrating embodied interaction with principled, data-informed design cycles, future iterations aim to push the boundaries of how learning engineering can support lifelike, ethical, and scalable practice for interpersonal competencies across education and professional domains.

References

- Arizona State University. (2022). Principled Innovation. <https://pi.education.asu.edu/>
- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*. <https://doi.org/10.1037/tmb0000058>
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.
- Cooperrider, D. L., & Srivastva, S. (1987). Appreciative inquiry in organizational life. *Research in Organizational Change and Development*, 1, 129-169
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale* (pp. 115-132). <https://doi.org/10.59668/2109.21735>

Fishbein, M., & Ajzen, I. (2010). Predicting and changing behavior: The reasoned action approach. Psychology Press.

Hiatt, J. M. (2006). ADKAR: A model for change in business, government, and our community. Prosci.

Mezirow, J. (1991). Transformative dimensions of adult learning. Jossey-Bass.

Rogers, C. R. (1959). A theory of therapy, personality, and interpersonal relationships, as developed in the client-centered framework. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 3, pp. 184–256). McGraw- Hill.

Totino, L., & Kessler, A. (2024). "Why did we do that?" A systematic approach to tracking decisions in the design and iteration of learning experiences. *The Journal of Applied Instructional Design*, 13(2). <https://doi.org/10.59668/1269.15630>



The Writing Analytics Tool: A Learning Engineering Approach to Designing AI-Supported Writing Instruction

Danielle S. McNamara, Andrew Potter, Zeinab Serhan, Manmeet Singh, Nishad Patne, Tracy Arner, Rod D. Roscoe, Laura K. Allen, & Scott A. Crossley

AI in Education

automated writing evaluation

learning engineering

Participatory Design

Writing Analytics

The Writing Analytics Toolkit (WAT) is an AI-driven platform designed to support writing instruction and research through transparent, theory-grounded writing analytics. Developed over six years through an IES-funded learning engineering initiative, WAT integrates natural language processing, machine learning, participatory design, and evidence-based instructional principles to serve students, teachers, and researchers within a single system. This paper describes WAT's architecture, analytics, and interfaces, and frames its development as a learning engineering case study. We document how stakeholder engagement, iterative design cycles, empirical validation, and scalability considerations shaped WAT's evolution. The work illustrates how learning engineering can guide the responsible design of AI-enabled educational tools that are pedagogically aligned, usable in authentic contexts, and extensible for research and innovation.

Introduction

Writing is a foundational academic and professional skill, yet it remains difficult to teach, assess, and study at scale (Graham, 2019). Large class sizes, limited instructional time, and the labor-intensive nature of writing feedback continue to constrain opportunities for deliberate practice and revision (e.g., Applebee & Langer, 2011). Automated writing evaluation (AWE) systems have emerged as a partial response, but many commercial AWE systems operate as opaque “black boxes,” offering scores without transparency or pedagogical grounding (McNamara & Potter, 2024; Wilson & MacArthur, 2024). In response to these challenges, the Writing Analytics Toolkit (WAT) was developed as an open, research-driven platform that provides actionable writing analytics to students, teachers, and researchers.

This paper has two goals. First, we describe WAT as it exists today, including its tools, analytics, and empirical foundations. Second, we frame WAT's development as a learning engineering process, illustrating how sustained stakeholder engagement, iterative design cycles, and evidence-centered analytic practices were used to align AI capabilities with instructional goals. In doing so, we position WAT as a concrete example of how learning engineering can guide the responsible development of AI-enabled educational technologies.

Learning Engineering as a Design Framework

Learning engineering is an interdisciplinary approach to designing and improving learning systems that integrates learning science, data-informed analytics, and human-centered engineering practices, emphasizing the translation of research insights into implemented, scalable learning solutions in authentic contexts (Baker et al., 2022; Goodell & Kolodner, 2023). Learning engineering is best understood as a process organized around defining a learning challenge and engaging in iterative cycles of creation, implementation, and investigation, in which learning solutions, analytics, and infrastructure are designed, deployed,

and examined using empirical evidence (Kessler et al., 2023). These cycles are iterative rather than linear, with evidence from each phase informing subsequent design decisions.

This process is particularly important for the development of AWE systems. Research findings on AWE suggest that the instructional value of these systems is contingent on meaningful integration into writing instruction and investment by writing instructors (Potter et al., 2025b; Wilson et al., 2025). Ongoing challenges in AWE include limited transparency, uneven instructional uptake, and tensions between automation and pedagogy. These challenges underscore the need for design approaches that foreground context and human decision-making (Shermis & Wilson, 2024). As AWE systems increasingly incorporate methods from artificial intelligence (AI), including large language models (LLMs), questions of human-centered design and evidence-informed implementation remain central (McNamara & Potter, 2024). These questions motivate the use of learning engineering as a framework for examining how writing technologies can be developed through iterative cycles of challenge definition, creation, implementation, and investigation to support effective, human-centered instruction at scale.

As a first step in applying the learning engineering process to WAT, we begin by defining the instructional and research challenges that motivated the development of the system and describing how the current versions of the system seek to address those challenges.

Defining the Challenge: Writing Instruction and Research at Scale

A central principle of the learning engineering process is the explicit definition of the learning challenge, which guides design decisions across subsequent cycles of creation, implementation, and investigation. In developing WAT, our research team identified two closely related challenges that persist in writing instruction and writing research: 1) supporting deliberate writing practice for students with meaningful feedback at scale, and 2) advancing the study of writing development through scalable, theory-informed analytic methods. As such, WAT was designed as a platform comprising two complementary systems: WAT Classroom (WAT-C) and WAT Researcher (WAT-R).

The current version of WAT-C is a web-based system that supports formative writing instruction for high school students, college students, and their instructors by integrating writing analytics into existing instructional workflows. Instructors use WAT-C to create writing tasks, select and configure analytics aligned with their instructional goals, and review student submissions across drafts. The instructor interface emphasizes control and flexibility, allowing teachers to determine which analytics are visible to students and how feedback is framed, thereby preserving instructors' roles as the primary evaluators of writing quality. Students interact with WAT-C through a dashboard that supports assignment management, review of analytic feedback, and submission of revised drafts. Across both interfaces, analytics are presented descriptively rather than evaluatively, with the goal of supporting interpretation, reflection, and revision rather than score optimization.

Whereas WAT-C is designed for instructional use in classrooms, WAT-R is intended for post hoc analyses of writing corpora. The current version of WAT-R is a natural language processing (NLP) tool that extracts linguistic features and component-level metrics derived from student writing. WAT-R is implemented as a standalone desktop application for macOS and Windows, designed to be usable without coding or technical setup. Researchers upload a corpora of student writing and use WAT-R to extract linguistic features and component-level metrics in a standardized format. These outputs are intended to support empirical analyses of writing development, instructional effects, and individual differences, as well as integration with statistical and machine learning workflows. WAT-R facilitates the development and testing of more advanced theoretical models of writing while promoting comparability across studies by reducing the time and technical expertise required for large-scale linguistic analysis.

In the sections that follow, we use the design and development of WAT-C as a case study to describe how learning engineering cycles of stakeholder engagement, analytic development, implementation, and investigation can be leveraged to develop human-centered technologies. Further, we discuss how these cycles continue to guide evaluation and refinement of the system's response to the persistent challenges in writing instruction and writing research.

Creation: Stakeholder Engagement through Participatory Design

Effective and sustainable classroom tools must align not only with theoretical models of learning but also with the practical, contextual realities of instruction. Instructional tools that are designed without sustained input from educators may fail to integrate into classroom workflows, may be misaligned with pedagogical goals, or introduce unintended burdens that limit adoption and impact. Therefore, considering stakeholders, particularly instructors, as active contributors during development is critical for ensuring usability, instructional relevance, and long-term sustainability. Participatory and co-design approaches support this alignment by foregrounding practitioner expertise, surfacing tacit instructional knowledge, and enabling design decisions that reflect authentic classroom constraints and opportunities.

Guided by these principles, the WAT-C development team purposely and systematically implemented a participatory research design (Wacnik et al., 2025) to develop WAT-C in collaboration with experienced writing instructors (see Li et al., 2022 for details). Secondary writing teachers were engaged as co-designers to ensure that early system decisions reflected instructional practice and classroom realities. In addition to teacher participants, the development team included researchers with expertise in psychology, applied linguistics, writing assessment, and learning analytics, as well as software engineers and UX designers. Several members of the research team also had prior experience designing automated writing evaluation systems, including the Writing Pal intelligent tutoring system (Roscoe et al., 2014).

Through this stakeholder-centered creation process, a central design requirement became clear: instructors did not want WAT-C to replicate or approximate predictive evaluation scores. Instead, they wanted analytics that could help them better understand patterns in students' writing and use that information to support instruction and revision. From a learning engineering perspective, this input prompted a design pivot away from formative scoring and toward descriptive writing analytics that characterize features of students' texts and support interpretation rather than judgment. These analytics were intended to function as sensemaking tools, enabling instructors to learn more about students' writing and enabling students to reflect on and revise their own work. Teachers further informed the design of system affordances emphasizing methods to communicate personalized feedback to their students. These stakeholder-driven decisions established the foundation for subsequent learning engineering cycles focused on analytic development, implementation, and empirical investigation.

Creation and Investigation: Iterative Analytics Development

Following the shift toward descriptive analytics, learning engineering cycles focused on developing analytic instrumentation that could characterize features of student writing at scale while remaining interpretable for instructional use. To accomplish this goal, WAT's analytic infrastructure integrates a set of established, open-source natural language processing tools from the [SALAT](#) (i.e., the Suite of Automatic Linguistic Analysis Tools, Crossley & Kyle, 2018). These tools, which have been validated and widely used in prior writing research, support the extraction of lexical, syntactic, and discourse-level features of student writing, including indices related to lexical sophistication and diversity, syntactic complexity, and local and global cohesion, as well as genre-relevant features for persuasive and source-based writing such as claims, citations, quotations, and indicators of source overlap (see Potter et al., 2025a).

Given the large number of linguistic features produced by this infrastructure, principal component analysis (PCA) modeling approaches were used to organize analytics into interpretable dimensions of writing that could be used in WAT-C to provide meaningful feedback to instructors and students. PCA was applied to hundreds of correlated features extracted from a corpus of persuasive essays, source-based essays, and summaries that were reduced into a smaller set of theoretically meaningful components, such as academic language use, elaboration, cohesion, and language variety (see Potter et al., under review). These components were validated against human judgments of writing quality and conceptualized as descriptive and facilitative representations of writing features rather than evaluative scores.

Subsequent studies refined and extended these analytics through iterative investigation. For source-based writing, a source integration construct was developed by modeling linguistic features related to citation practices, quotation, plagiarism

indicators, and semantic overlap with source texts, with evidence of strong alignment with human ratings and generalizability across prompts and datasets (Potter et al., under review). Related work developed and validated a global cohesion metric and extended its use to evaluate revisions generated by LLMs, illustrating how validated linguistic analytics can be applied to emerging instructional contexts while maintaining construct validity (Potter et al., in press b).

Implementation and Investigation: System Refinement and Scaling

Following the development of WAT's writing analytics, subsequent learning engineering cycles focused on extended investigation and system refinement to assess feasibility, usability, and instructional relevance prior to broader deployment. A second participatory design study in collaboration with writing instructors examined how WAT's analytics, workflows, and configuration options aligned with instructional needs (see Potter et al., in press a). Instructor feedback informed targeted refinements to metric presentation and customization features, as well as principled decisions about scope. Notably, instructors reported limited instructional use of summaries, reflecting shifts in writing practice in the context of generative AI tools. In combination with emerging evidence that summaries may offer weaker support for writing development relative to other genres (McNamara et al., 2024), these findings motivated the removal of summaries from the instructional interface in favor of deeper support for persuasive and source-based writing.

In parallel, system-level refinements addressed scalability and readiness for deployment. WAT's infrastructure was re-architected to improve computational efficiency and support future integration with learning management systems through secure cloud-based services and selective index computation. The cloud infrastructure was designed and tested to securely manage user data, support concurrent access by multiple users, and scale dynamically under increased load, ensuring reliability and data protection as adoption increases. These changes were designed to enable real-time student feedback while minimizing instructional and technical burden.

Building on these refinements, the next learning engineering cycle focuses on integrating generative AI to support students' interpretation and use of writing analytics during drafting and revision. Instructor feedback from the participatory design study indicated that students increasingly expect personalized, conversational support aligned directly with their own writing, and that instructors would be hesitant to adopt analytics tools that do not incorporate such affordances. In response, ongoing work explores the integration of a chatbot powered by a large language model to scaffold writing and revision using WAT-produced analytics, while simultaneously maintaining human-in-the-loop control over feedback and evaluation. Planned experimental studies will examine how students interact with this chatbot and evaluate its effects on revision quality, writing outcomes, and instructional use, extending WAT's learning engineering cycle through systematic investigation of this next refinement.

Conclusion

The Writing Analytics Toolkit case study illustrates how learning engineering can guide the development of AI-enabled educational tools over extended timelines. Across its development, learning engineering principles shaped decisions about analytic representation, system architecture, and refinement, resulting in a platform that prioritizes interpretability, instructional relevance, and usability alongside scalability.

As both writing instruction and educational research increasingly intersect with advances in artificial intelligence, WAT provides an example of how AI-supported tools can be designed to augment human judgment and practice opportunities in educational contexts. More broadly, this work underscores the value of learning engineering as a framework for ensuring that emerging educational technologies remain aligned with evidence-based instructional practices and user needs. Continued

investigation of WAT and its future iterations will further inform how learning-engineered tools can responsibly support teaching and learning.

Acknowledgments

This work was supported by the Institute of Education Sciences (R305A180261) as well as the Learning Engineering Institute at Arizona State University. We thank the National Writing Project teachers, instructors, students, and research team members who contributed to WAT's development.

References

- Applebee, A. N., & Langer, J. A. (2011). A snapshot of writing instruction in middle schools and high schools. *English Journal*, 100(6), 14–27. <https://doi.org/10.58680/ej201116413>
- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*, 3, 1–23. <https://doi.org/10.1037/tmb0000058>
- Crossley, S. A., & Kyle, K. (2018). Assessing writing with the tool for the automatic analysis of lexical sophistication (TAALES). *Assessing Writing*, 38 (1), 46-50.
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. *Journal of Military Learning, Conference Edition*, 46–59.
- Goodell, J., Kolodner, J. L., & Kessler, A. (2023). Learning engineering applies the learning sciences. In J. Goodell & J. L. Kolodner (Eds.), *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond*. Routledge. <https://doi.org/10.4324/9781003276579>
- Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*, 43(1), 277–303. <https://doi.org/10.3102/0091732X18821125>
- Kessler, A., Craig, S. D., Goodell, J., Kurzweil, D., & Greenwald, S. W. (2023). Learning engineering is a process. In J. Goodell & J. L. Kolodner (Eds.), *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond*. Routledge. <https://doi.org/10.4324/9781003276579>
- McNamara, D. S., & Potter, A. (2024). The two U's in the future of automated essay evaluation: Universal access and user-centered design. In *The Routledge international handbook of automated essay evaluation* (pp. 590–608). Routledge.
- McNamara, D. S., Watanabe, M., Huynh, L., McCarthy, K. S., Allen, L. K., & Magliano, J. P. (2024). Summarizing versus rereading multiple documents. *Contemporary Educational Psychology*, 76, 102238. <https://doi.org/10.1016/j.cedpsych.2023.102238>
- Potter, A. H., Crossley, S. A., Öncel, P., Allen, L. A., Xu, X., Balyan, R., Arner, T., Roscoe, R. D., & McNamara, D. S. (manuscript under review). From words to components: Mapping linguistic features in source-based writing.
- Potter, A. H., Serhan, Z., Patne, N. A., Öncel, P., Ahmed, I., Arner, T., Islam, R., Allen, L. A., Roscoe, R. D., Crossley, S. A., & McNamara, D. S. (in press a). Human-AI Collaboration in Participatory Design: Refining the Writing Analytics Tool. *Journal of Educational Data Mining*.

- Potter, A., Shortt, M., Goldshtein, M., & Roscoe, R. D. (2025a). Assessing academic language in tenth-grade essays using natural language processing. *Assessing Writing*, 64, 100921. <https://doi.org/10.1016/j.asw.2025.100921>
- Potter, A., Tian, Y., Balyan, R., Watanabe, M., & McNamara, D. S. (in press b). Revising with generative AI: Supporting writing instruction through validated cohesion metrics. In Z. Philippakos (Ed.), *AI literacy: Foundations, practices, and innovations in education*. Routledge.
- Potter, A., Wilson, J., Roscoe, R.D., Arner, T., & McNamara, D.S. (2025b). Computer-based writing instruction. In C.A. MacArthur, S. Graham, & J. Fitzgerald (Eds.). *Handbook of writing research* (3rd ed.). Guilford Press.
- Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, 34, 39-59. <https://doi.org/10.1016/j.compcom.2014.09.002>
- Shermis, M. D., & Wilson, J. (2024). Introduction to automated essay evaluation. In *The Routledge international handbook of automated essay evaluation* (pp. 3–22). Routledge.
- Thai, K. P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. L. (2023). Learning engineering is human-centered. In J. Goodell & J. L. Kolodner (Eds.), *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond*. Routledge. <https://doi.org/10.4324/9781003276579>
- Wilson, J., & MacArthur, C. (2024). Exploring the role of automated writing evaluation as a formative assessment tool supporting self-regulated learning in writing. In *The Routledge international handbook of automated essay evaluation* (pp. 197–220). Routledge.
- Wilson, J., Palermo, C., Myers, M. C., Cruz Cordero, T., Eacker, H., Potter, A., Coles, J., May, H., & Zhang, F. (2025). Impact of automated writing evaluation (AWE) on middle school students' writing outcomes. *American Educational Research Journal*. <https://doi.org/10.3102/00028312251388655>



Towards Automated Detection of Struggling Student Programmers

Sanjita Patwardhan & Ruben Acuña

Educational Data Mining

Exploratory Data Analysis

Struggle

In programming courses, it is often difficult for instructors to identify students who struggle while coding. Fortunately, automated assessment tools used in courses provide a way to capture data about programming activity. Using this data source provides the foundation for developing a machine learning model to automatically classify students who are struggling, even in large courses. Such a model would help instructors target interventions to help struggling students. In this paper, we provide a step towards creating this model. We have conducted preliminary work focused on identifying features with the potential to indicate struggle, performing feature engineering to extract them, and then conducting an exploratory data analysis on real data to visualize outliers and assess feasibility.

Introduction

Identifying struggling students in computing courses remains a persistent challenge for many instructors. Students working on assignments at home are unobserved by an instructor, and even those completing an in-class activity may not catch the instructor's attention in a large class. However, courses often make use of automated assessment tools (AATs), which allow students to submit assignments and receive feedback on which parts of their assignments have correct functionality by executing a set of instructor-developed tests (Messer et al., 2024). AATs let us implicitly observe students by capturing development activity, such as submission attempts, compilation status, and progress according to test cases. There is extensive work on educational data mining (see Inhantola et al., 2015), with prior research on block-based languages (BBLs) showing how to detect struggling learners and time interventions (e.g., Dong et al., 2021; Tabarsi et al., 2022). However, much remains unknown about which patterns are most indicative of struggle and how they are shown in real datasets. When instructors can identify struggling students earlier, they are able to direct support, enhance activities, and help prevent disengagement. Without automated detection, many patterns of process struggle remain hidden unless reported by students. Unlike work with BBLs, we examine data for a general-purpose language (Java), requiring feature engineering before conducting exploratory data analysis (EDA) on submission logs. Our guiding research question is: can the features in student code traces predict whether a student will get stuck on an assignment? In this preliminary work, we investigate features such as the number of submissions each student makes, the coding time gaps, and temporal patterns of work. This work was structured as a nested Learning Engineering cycle that focused on identifying relevant data features within the challenge stage of our project to identify struggling students.

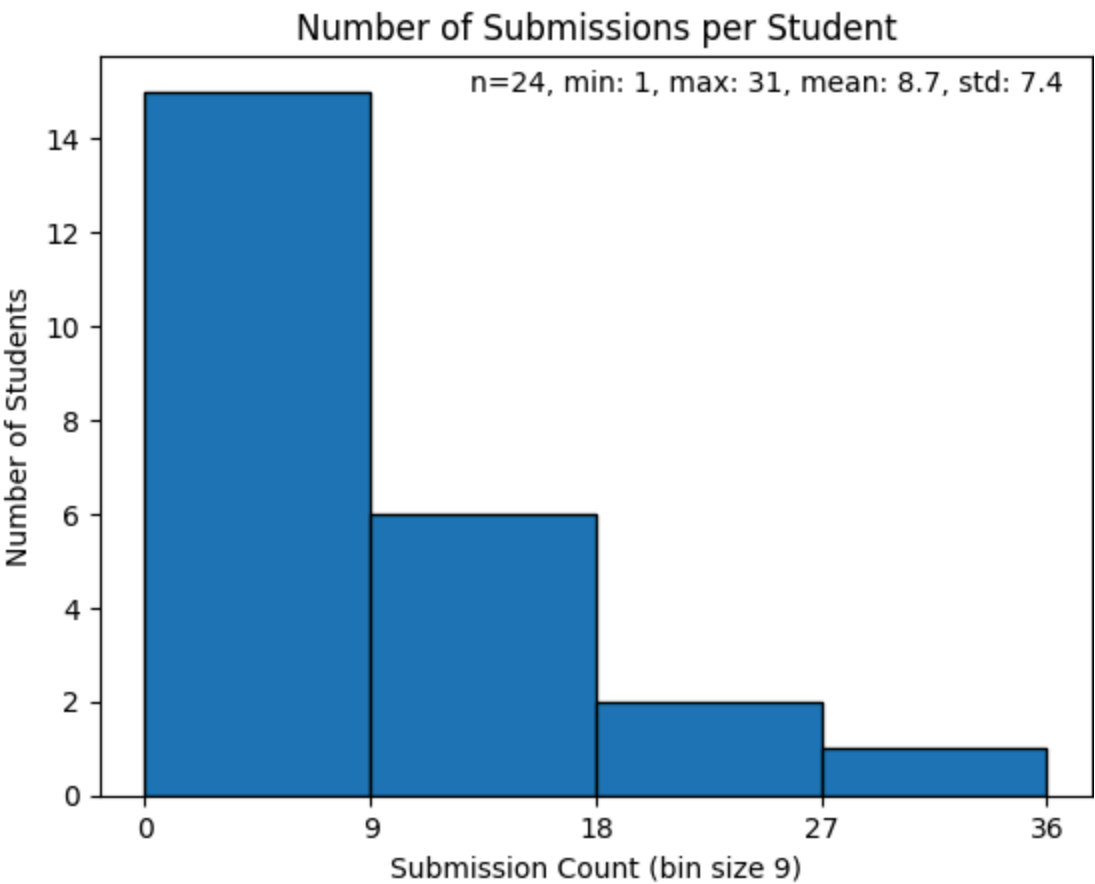
Methods

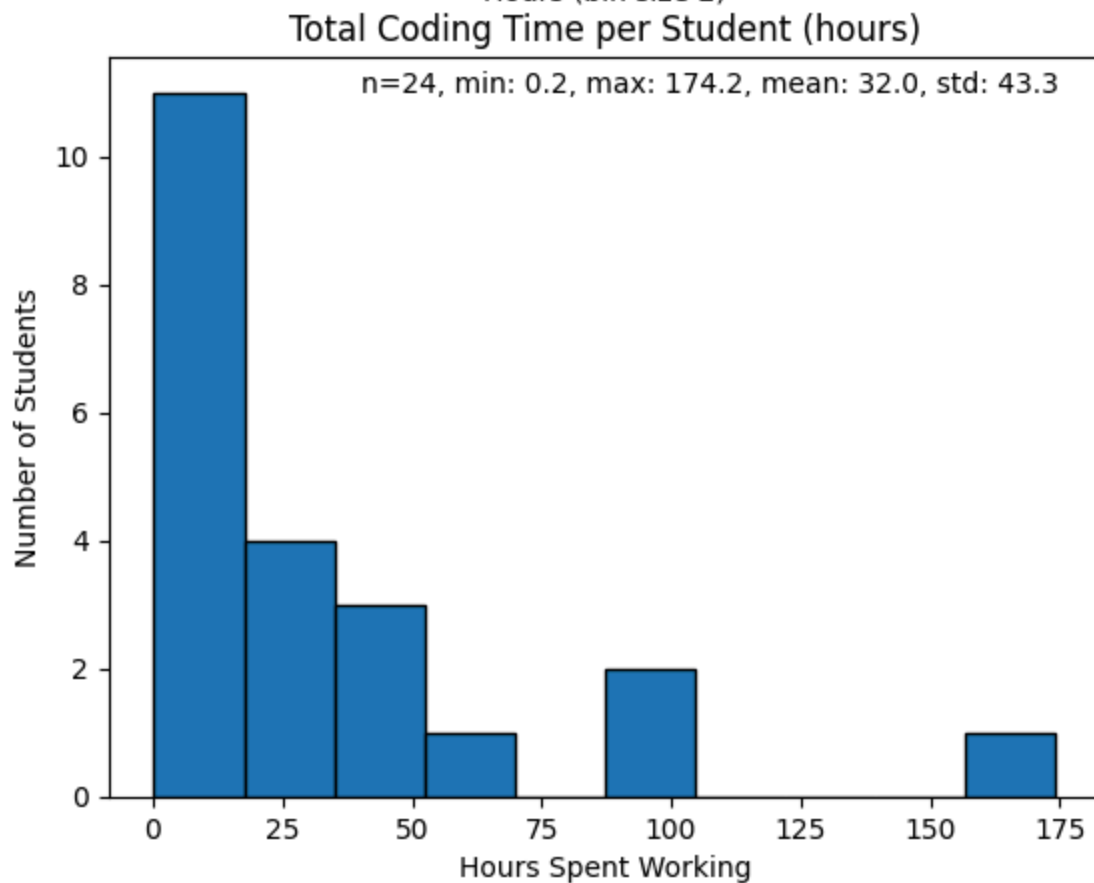
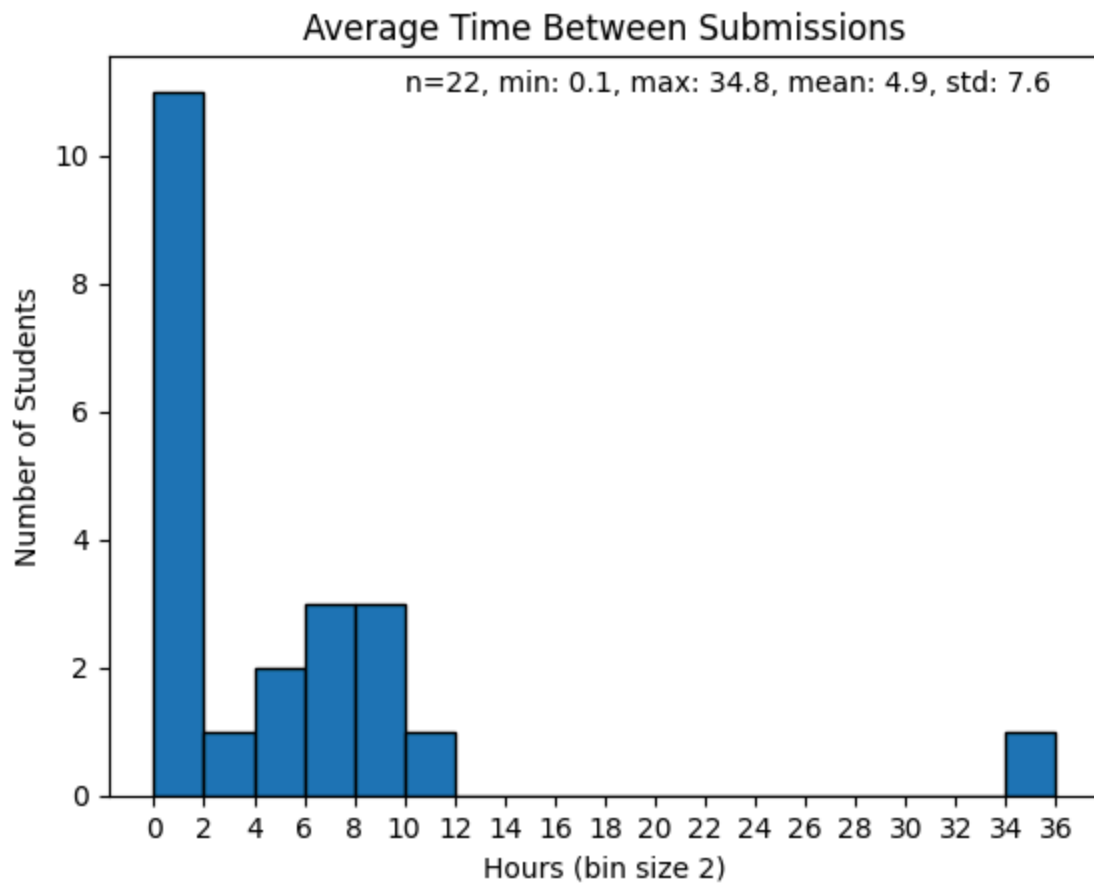
To investigate which features may signal student struggle, we implemented a data analysis pipeline that extracts features from Gradescope logs. Our script processes students' timestamped submission history and computes measures that reflect coding behaviors. We focused on four features that represent student submission behavior: 1) Number of Submissions: The number of times each student submitted during the assignment. 2) Time Between Submissions: For each student, we compute the time between consecutive submissions. The longer gaps may indicate debugging times or a break. 3) Time of Day of Work: Using the attempt timestamps, we determine the hour(s) in which work occurred. This helps reveal working patterns, such as last-minute bursts. 4) Total Coding Time: We calculate the delta between a student's first and last submission. This approximates a lower bound for the time each student engaged with the assignment as seen by AAT activity. These features were selected because we have informally observed students repeatedly submitting code to ensure it matches the assignment requirements.

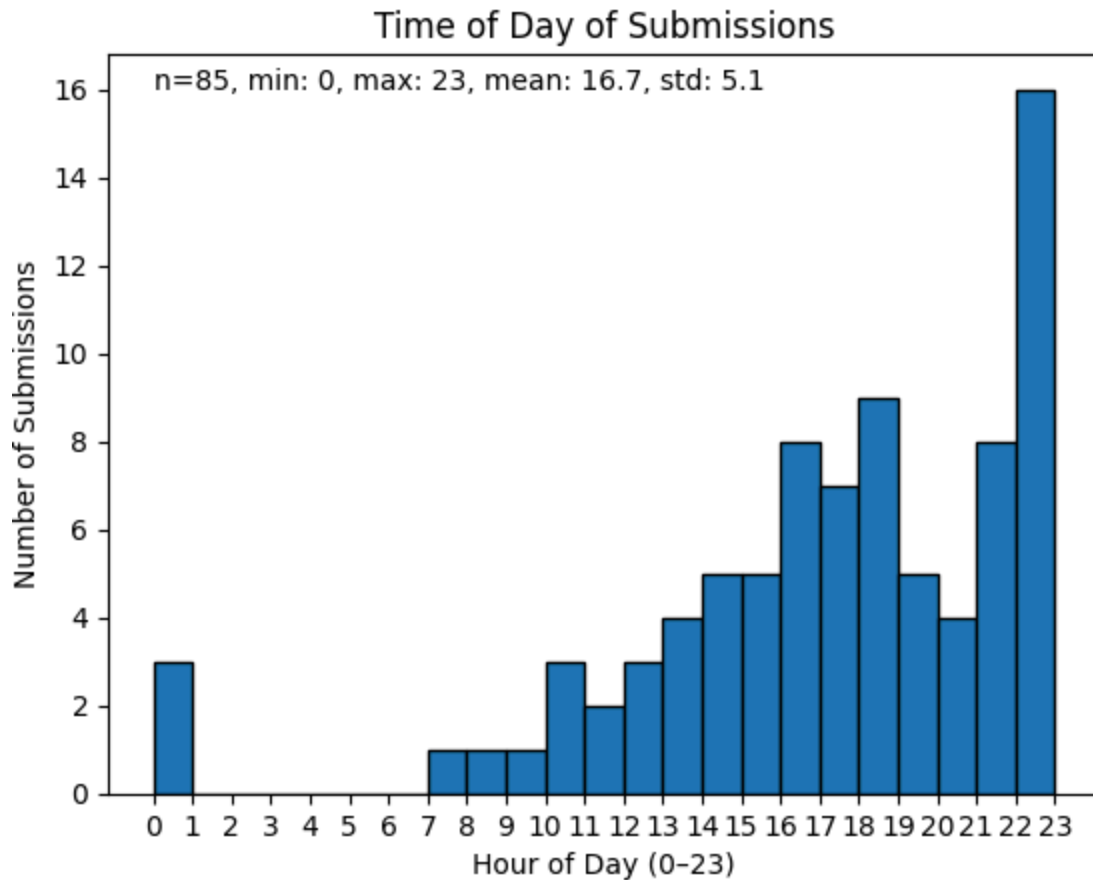
Results

We applied our methodology to data from a face-to-face sophomore-level course on data structures & algorithms (described in Acuña & Bansal, 2024) taught in the spring of 2025. The course is taught in Java and uses an AAT to assess assignments on topics such as lists, trees, hash tables, and graphs. This paper focuses on a warm-up assignment in which students implemented a matrix class (n=24). Histograms for the features are shown in Figure 1.

Figure. 1. Visualizations for four extracted features on data structures & algorithm dataset.







These plots indicate several trends. Submission counts ranged from single attempts to 31, reflecting different strategies from students and how they went about completing the assignment. Average time between submissions displayed more students working on the coding assignment with few breaks in between, with most students submitting frequently and a small number exhibiting long pauses that may indicate extended debugging. Total coding time also varied widely, with some students completing work quickly while others engaged across longer multi-hour spans. Time-of-day patterns showed clustering around late afternoon and evening hours, with many submitting a couple of hours before the 11:59 pm deadline. Based on the extracted features (and others), we plan to develop a machine learning classifier.

Discussion

Our findings suggest features that may meaningfully reflect students' coding processes. The distributions we observed, especially in submission counts, timing gaps, and total coding time, highlight that students do not engage with the assignment uniformly. Some patterns, such as long debugging pauses, could indicate difficulty resolving errors, as seen by Dong et al. (2021). Other students exhibited short coding durations and few submissions, suggesting efficient completion or disengagement. One limitation is the varying sample size across features. Time between submissions could not be calculated for every student, since it requires at least two submissions. Additionally, time-of-day patterns, while useful, may reflect external factors such as schedules rather than factors of struggling. Despite this, our EDA demonstrates that submission-level behaviors contain measurable structure that could inform models. These extracted features appear interpretable, making them candidates.

Future work will focus on building an explainable machine-learning model using the engineered features identified in this analysis. By utilizing an explainable model, we will help instructors understand why students are flagged as struggling and provide insight into which features indicate struggle. Expanding the dataset and validating against instructor-identified cases of struggle will strengthen the computer system's reliability. Ultimately, our goal is to design an early-warning tool to increase pass rates for the computing courses.

References

- Acuña, R., & Bansal, A. (2024). WIP: Characterizing Student Programming Activity. 2024 IEEE Frontiers in Education Conference (FIE), 1–5.
- Dong, Y., Marwan, S., Shabrina, P., Price, T., & Barnes, T. (2021). Using Student Trace Logs To Determine Meaningful Progress and Struggle During Programming Problem Solving. Proceedings of The 14th International Conference on Educational Data Mining (EDM).
- Ihantola, P., Vihavainen, A., Ahadi, A., Butler, M., Börstler, J., Edwards, S. H., Isohanni, E., Korhonen, A., Petersen, A., Rivers, K., Rubio, M. Á., Sheard, J., Skupas, B., Spacco, J., Szabo, C., & Toll, D. (2015). Educational Data Mining and Learning Analytics in Programming: Literature Review and Case Studies. Proceedings of the 2015 ITiCSE on Working Group Reports, 41–63.
- Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2024). Automated Grading and Feedback Tools for Programming Education: A Systematic Review. ACM Transactions on Computing Education, 24(1), 1–43.
- Tabarsi, B. T., Limke, A., Reichert, H., Qualls, R., Price, T., Martens, C., & Barnes, T. (2022). How to Catch Novice Programmers' Struggles: Detecting Moments of Struggle in Open-Ended Block-Based Programming Projects using Trace Log Data.



User experience design of AI-assisted human-technology ecosystem for writing assessment

Li Liang

academic integrity

Generative AI

GPTZERO

Human IA (intelligence augmentation)

User Experience Design

Writing Assessment

Generative AI (GenAI) raises increasing concerns as it has been swiftly used for academic misconduct and the tech communities have not figured out effective solution to detect AI-written text, particularly in the context of writing assessment. This threat to academic integrity and assessment validity urgently calls for effective change in the user experience (UX) design of writing assessment before we can detect AI-written text one day. How to build up a sustainable and valid GenAI-assisted human-technology ecosystem for writing assessment with academic integrity? What practical goals of writing assessment can we use GenAI agents to meet and extend to its full potential for human IA (intelligence augmentation)? This experimental pilot study reviews the user experience design of four GenAI tools and GPTZERO with socio-cultural approach to explore these questions from user-centered perspective. Three principles for user experience design of AI-assisted writing assessment are synthesized for mastery learning.

Introduction

Almost every time when a record-breaking new technology came into being in the human history, be it electricity over candles, or cars over horses, we users were confronted with the similar process of shock, doubt, rejection, fight, taste, confirmation, accept, like, prevalent vogue everywhere, and so on. The similar story happens with generative AI (GenAI) tools such as ChatGPT, Claude, Grok and Gemini. On the other hand, GenAI raises increasing concerns as it has been swiftly used for academic misconduct and the tech communities have not yet figured out effective solution to detect AI-written text, particularly in the context of writing evaluation or assessment. For example, the International Conference on Machine Learning (ICML), has officially forbid researchers from using AI tools like ChatGPT to write scientific papers. Though GPTZERO can detect most of AI-written work, it cannot when some words in the AI-written work are changed by humans.

Given such facts mentioned above, how can we effectively and validly design our user experience (UX) of GenAI-assisted human-technology ecosystem for the writing assessment? What practical goals of writing assessment can we use GenAI agents to meet and extend to its full potential for human IA (intelligence augmentation)?

Methods

This experimental pilot study conducts a small-scale experiment with a sample of four GenAI tools— ChatGPT, Claude, Grok and Gemini, and one AI-writing detection tool GPTZERO, to explore and review the UX design of using GenAI tools and GPTZERO in writing assessment with socio-cultural approach. The sampling details, prompts, interaction logs as empirical evidence from which key conceptual principles were derived in this paper are illustrated in GitHub repository <https://github.com/llcfy2023/User-experience-design-of-GenAI-assisted-human-technology-ecosystem-for-writing-assessment/tree/main>. Content analysis in socio-cultural study was adopted for the data analysis.

Results

Based on the human-centered nested learning engineering cycle (Goodell, Kessler & Schatz, 2023; Craig et al., 2025), socio-cultural theory (Vygotsky, 1986; Lave & Wenger, 1991) and modern assessment theory (Andrich & Marais, 2019; Tognolini & Stanley, 2007), three principles for user experience design for AI-assisted writing assessment are synthesized for UX design of mastery learning: socio-cultural fitness, incremental validity, and empower human IA for all people. A new writing assessment mode of purpose is then proposed for the front-line educators in GenAI age.

Discussion

1. socio-cultural fitness

Undeniably, the digital technology like internet and GenAI not only expands our knowledge/information to a historically unprecedented level, but also brings us new challenges and enhances our ability to do things at an unheard-of new level ever. We cannot ban internet from our students' life in the internet age, nor can we forbid their use of GenAI in AI age. Mathematician Dr. Terence Tao in UCLA noted in his 2022 mathstodon post that it is of no use to fight against the new tech in the long term and educators need to adopt an "open books, open AI" mode of assessment to adapt to the times.

The search engines like google can offer us information on our questions, but what makes GenAI more intelligent than the search engines are that it fits more with human culture in intelligent interactions and enquiries, not just at the interactional level of accessing knowledge or information as google, but also at the interactional level of more or less synthesized ideas or thinking. This design path based on socio-cultural fitness is obviously different from the conventional path based on neuro-mimic design. The better fit with the human culture of interaction that AI tools can have, the more human-like intelligent and instant popularized they will be. An AI tool that fits well with human culture and can provide us the same assistant service that we need for our life in human society, can be another successful way besides the neuro-mimic AI development to realize AI design to its record-breaking level.

With the nature of socio-cultural fitness, the goal of UX design like ChatGPT and GPTZERO is to empower the human culture of interaction and communication in learning as a collaborative intelligent assistant in response to the people's prompts, not to replace people in the division of labor in the digital activity of GenAI age. The principle of socio-cultural fitness also offers UX design in GenAI an effective and simple means to avoid ethical pitfalls that may potentially threaten or harm the healthy cultural norms of the human society.

Another outstanding benefit of this socio-cultural fitness in UX design is that it can greatly enhance the UX value of usability and desirability. How long will it take you to learn to use ChatGPT or GPTZERO? Often only within a couple of seconds. This is because the system interacts just like a real person, providing us a genuine natural human communication experience of a chat that we have already been very familiar with.

1. incremental validity

Incremental validity is the second key principle for effective UX design of GenAI tools in the context of writing tasks. Its natural dialogue to exchange ideas fits well for providing feedback, answering questions in formative assessment (FA) as intelligent agents in learning and assessment. But a major reason for the low acceptability in some cases of evaluation or even a ban on ChatGPT in New York public schools is that the UX design of ChatGPT omitted any technical measures to increase the evidence-based validity of its answers in the users' evaluation, nor provide specific sources of reference to the users for more trustworthy human-technology interaction and collaboration. In contrast, the UX design of GPTZERO pinpoints exactly this principle of incremental validity: GPTZERO clarifies the evidence-based measures and evaluative framework that the intelligent analysis outcome is based on. This undoubtedly enhances the transparency of its background operation and increases the trustworthiness in its users.

2. Empower human IA for all learners via GenAI

If AI tools are designed for the coding people only, then that means a majority of the population will be left out, which is in opposite to our goal for an inclusive and sustainable learning engineering. This principle serves to structure the AI-assisted human-technology ecosystem as a user-centered inclusive and sustainable one for learning and assessment. The change unavoidably leads to a change in what values in the workforce market: those who are excellent in collaboration with AI tools to solve problems and complete tasks with good efficiency at work will be in great demand in the near future. This new change in the values of collaborative capacity with AI tools in the workforce market will of course prompt the need for a new assessment mode for the students using AI agents, especially for the writing assessment.

3. A new writing assessment mode of purpose

This new mode will be discussed via two key questions: what to assess and how to assess. Before figuring out these two questions, we'll need to be clear of the new purpose of the assessment mode for the students using AI tools. As discussed above, the purpose of the new assessment mode is not to forbid the students from using AI tools, but to assess how well the students complete the assessment task in collaboration with AI tools such as ChatGPT and GPTZERO.

Firstly, we will have to redefine what to assess in the writing assessment. With the open access to internet and availability of AI tools to all students such as ChatGPT, it is of somewhat limited value to assess the students' knowledge and basic thinking skills such as memory anymore, since everyone can get equal access to internet for all human knowledge over the thousands of years in a second. What matters more than merely knowledge and basic thinking skills in the digital age, is the students' ability in problem-solving, higher-order thinking, collaboration with other students and AI tools, value of the culture and society, etc.

According to the structure of the activity theoretical framework (Leontiev, 1978), it is also necessary to consider the new change in the division of labor in the digital activity of learning: the students will shoulder more responsibility as independent leaders, managers or principal investigators in their collaboration with AI tools, different from their traditional role of somewhat passive object to follow the teachers' leadership. Thus, the students' leadership and organization ability in their collaboration with AI tools in accomplishing the assessment tasks may also be included in the content of the writing assessment in university.

Another key question is how to assess the students using AI tools in the context of writing assessment in university. The teachers who would like to give their students writing assessment tasks will need to refresh their repertoire of assessment means and strategies according to their specific assessment purposes. AI-assisted human-technology ecosystem works effectively where speed, encyclopedic knowledge, or precision are key for a writing assessment task. For example, GenAI provides a short-cut to the students who need to complete a writing task but don't have sufficient time. If the teachers want to assess how well the students can collaborate with GenAI tools to complete a writing assessment task in a short time, then the teachers may design a writing assessment task that requires the students to finish in collaboration with AI tools within a very short period of time. Last but not least, teachers also need to be careful in their wording in the writing assessment task description so as to establish the identity of AI tech as assistant tools only, not as another robotic agent beings for the task.

References

ChatGPT Homepage, <https://chat.openai.com/auth/login>, last accessed 2025/10/15.

Claude Homepage, <https://claude.ai/chat>, last accessed 2025/10/15.

Gemini Homepage, <https://gemini.google.com/app>, last accessed 2025/10/15.

GPTZERO Homepage, [http://www.GPTZERO \[2\].me](http://www.GPTZERO [2].me), last accessed 2025/10/15.

Grok Homepage, <https://www.grok.com>, last accessed 2025/10/15.

Andrich, D., & Marais, I. (2019). A course in Rasch measurement theory: Measuring in the educational, social and health sciences. Springer Singapore.

Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). <https://doi.org/10.59668/2109.21735>

Goodell, J., Kessler, A., & Schatz, S. (2023). Learning Engineering at a Glance. Journal of Military Learning, <https://www.armyupress.army.mil/Journals/Journal-of-Military-Learning/Journal-of-Military-Learning-Archives/Conference-Edition-2023-Journal-of-Military-Learning/Engineering-at-a-Glance/>

Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge University Press.

Tao, T. (2022). AI tools like #ChatGPT will soon be capable of answering a large fraction of traditional university homework-type questions [Social media post]. Mathstodon. <https://mathstodon.xyz/@tao/109543141003492779>

Tognolini, J., & Stanley, G. (2007). Standards-based assessment: a tool and means to the development of human capital and capacity building in education. The Australian Journal of Education, 51(2), 129-145. <https://doi.org/10.1177/000494410705100203>

Vygotsky, L. S. (1986). Thought and language. MIT Press.



Using AI to Bridge Technology Gaps in Higher Education

Universities rely on enterprise technology platforms to deliver essential services, yet vendor-provided systems often impose rigid constraints that limit institutional responsiveness to pedagogical needs. This paper presents a learning engineering approach to bridging technology gaps through AI-augmented local solutions. Following the learning engineering (LE) process (Goodell et al., 2023), we addressed a structural limitation in our survey platform that prevented longitudinal analysis of student self-assessments. Through human-centered stakeholder analysis (Schatz et al., 2022; Thai et al., 2022), we identified requirements for visualizing individual student growth on Program Learning Objectives (PLOs) over time. We then leveraged large language models to develop a FERPA-compliant, locally-run pipeline that transforms fragmented course-level survey data into longitudinal student trajectories. Our contribution is twofold. First, we demonstrate how institutions can create adaptive layers within existing ecosystems that address context-specific needs and function as tech augmenters instead of passive consumers. Second, we provide practitioners with a framework guiding them through friction identification, feasibility assessment, process decomposition, and implementation criteria including data governance, maintainability, and scalability. Implementation revealed three key affordances: (1) faculty gained visibility into individual learning trajectories

previously obscured by course-level aggregates, (2) curricular gaps became evident through cross-course PLO progression mapping, and (3) accreditation documentation improved through evidence-based growth narratives. This work extends learning engineering practice by demonstrating how readily available AI tools can enable local innovation within complex institutional technology environments, offering a replicable approach for institutions facing similar constraints.

Pulkit Goyal

Arizona State University; puhlkit@gmail.com

Problem and Context

Universities depend on vendor-provided systems that often impose rigid, one-size-fits-all constraints to deliver essential student services. When limitations emerge in these platforms, institutions typically wait months or years for updates, leaving critical needs unmet. This challenge is particularly acute in program assessment, where faculty and administrators need longitudinal views of student development but existing survey platforms lack the capability to aggregate individual responses across multiple administrations.

At Arizona State University, we encountered this friction point with our institutionally-supported survey platform (QuestionPro). Faculty used end-of-course self-evaluations to track students' perceived growth on Program Learning Objectives (PLOs) across an academic program. However, neither QuestionPro nor its predecessor (Qualtrics) could display a student's longitudinal progression across all surveys—a fundamental requirement for understanding individual learning trajectories and evaluating curricular effectiveness. This structural limitation prevented evidence-based program improvement and complicated accreditation documentation.

This challenge illustrates a broader learning engineering concern: How can institutions close gaps in their technology ecosystems without waiting for external development cycles while ensuring solutions remain compliant, scalable, and grounded in pedagogical needs?

Innovation and Approach

Rather than accepting these limitations, we adopted a learning engineering framework (Goodell et al., 2022) to systematically address this challenge through an AI-augmented solution. Following the learning engineering process—Challenge, Create, Implement, Investigate—we developed a locally-run data processing pipeline that transformed fragmented survey data into actionable longitudinal insights.

Challenge Phase: Through stakeholder interviews with faculty and administrators, we identified the core need: visualizing individual student growth trajectories across PLOs throughout their program. We conducted a needs analysis following human-centered design principles (Schatz et al., 2022), documenting specific analytical requirements and constraints including FERPA compliance and the need for offline operation.

Create Phase: We leveraged AI tools (specifically, ChatGPT) to develop a custom analytical pipeline with three components: (1) CSVs extracted from QuestionPro, (2) AI-coded preprocessing scripts to merge and normalize surveys, and (3) a search interface enabling faculty to query individual students and generate longitudinal visualizations. We documented each design decision, its evidence base, and anticipated outcomes to maintain transparency and support iteration.

The innovation lies not in replacing vendor platforms but in creating an "adaptive layer" within the existing technology stack. This approach exemplifies learning engineering's human-centered focus: we designed from the perspective of faculty and administrators who needed these insights, not from what the technology vendor prioritized.

Implementation: All processing occurs locally without internet connectivity, ensuring FERPA compliance and institutional data governance standards. The solution operates as a supplementary tool that faculty can use alongside existing survey infrastructure.

Early Insights and Expected Outcomes

Initial deployment revealed three key affordances:

1. **Individual Learning Trajectories:** Faculty can now visualize how each student's self-assessed competence evolves across courses, revealing patterns previously invisible in course-level aggregates.
2. **Curricular Alignment:** By mapping PLO progression across the program sequence, we identified courses where students showed minimal growth on relevant outcomes, signaling potential redesign needs.
3. **Accreditation Evidence:** The compiled longitudinal data creates clear, evidence-based narratives of student development over time, directly supporting program review and accreditation requirements.

Beyond these analytical benefits, the project demonstrates a methodological shift. Rather than positioning instructional designers and faculty as passive consumers of vendor-provided functionality, the learning engineering approach augmented by accessible AI tools enables them to become active technology augmenters. This represents what (Thai et al., 2022) describe as applying learning sciences using "human-centered engineering design methodologies and data-informed decision-making."

Implications for Learning Engineering

This work contributes to learning engineering practice in three ways:

Practical Framework: We developed a lightweight decision framework for identifying when AI can overcome technology stack limitations: (1) Define the friction point in existing tools, (2) Research feasibility through process decomposition, (3) Map the exact procedural steps required, and (4) Use AI to automate the mapped process. This framework helps teams distinguish between solvable gaps and those requiring vendor-level solutions.

Sociotechnical Design: Our approach exemplifies how learning engineering solutions must consider the interplay between technology capabilities and human workflows. By engaging stakeholders through interviews and needs analysis, we ensured the tool fit naturally into existing practices rather than imposing new procedural burdens.

Scalable Local Innovation: By demonstrating how institutions can layer compliant, locally-developed intelligence onto legacy systems, we offer a model for closing capability gaps without waiting for enterprise vendors. This approach is particularly valuable in higher education, where diverse institutional contexts require customization that enterprise solutions cannot provide.

Next Steps: We are instrumenting the tool to collect usage data and plan to investigate its impact on faculty decision-making and program improvement cycles. We seek collaborations with other institutions facing similar technology gaps to test the framework's generalizability.

References

- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. Army University Press.
<https://www.armyupress.army.mil/Journals/Journal-of-Military-Learning/Journal-of-Military-Learning-Archives/Conference-Edition-2023-Journal-of-Military-Learning/Engineering-at-a-Glance/>
- Schatz, S., Thai, K.-P., Craig, S. D., Schoenherr, J. R., Lis, J., & Kolodner, J. (2022). Human-centered design tools. Learning Engineering Toolkit, 279–301. <https://doi.org/10.4324/9781003276579-17>
- Thai, K.-P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. (2022). Learning engineering is human-centered. Learning Engineering Toolkit, 83–123. <https://doi.org/10.4324/9781003276579-7>



Educational Technology Informed by Learning Engineering

[CoreTrust as a Civic Learning Tool: Community Capital for Inclusive Digital Governance](#)

[Engineering a Semantic Topicality Instrument for Multiple-Choice Question Quality Control](#)

[Design and Development of Two Digital Solutions for Promoting Harmonious Relationships: Buddy Up and Squeez](#)

[From Immersion to Action: How VR Influences Behavioral Intentions Around Advanced Water Purification](#)

[From Immersion to Behavioral Intentions: Virtual Reality and Advanced Water Purification](#)

[How Students Really Use Courseware: Visualizing Student Pathways in Integrated Chemistry Courseware](#)

[REAL CHEM Action Research Through the LearnLab Summer School](#)

[SCALED: A Transferable Learning-Engineering Model for Large-enrollment STEM Courses](#)

[Terracotta: Lowering Barriers to Experimental Education Research](#)

[The Education Tree: A New Theoretical Model for P-20 Education and Development](#)

[The Impact on Cognition and Motivation Using Gaming, Simulation, and Visual Learning in Military Flight Training](#)

[The Promise of Scenario-Based Assessments for College Instruction](#)

[“Walk It Out”: An Embodied and Mobile AI Tutor for STEM Education](#)

[Impact of Varying the Playback Speeds of Educational Content on Learning and Engagement](#)



CoreTrust as a Civic Learning Tool: Community Capital for Inclusive Digital Governance

Mark Roseland & Mozhgan Ansari

Civic Learning

Community Capital Compass

Digital Governance

This paper explores CoreTrust as a civic learning tool that helps practitioners and educators make sense of complex community data through a community capital lens. The study examines how CoreTrust can turn existing data into a shared basis for discussion and decision-making. We drew on an ongoing pilot in Casa Grande, Arizona, using a qualitative design informed by participatory action research. Data includes planning documents, platform artifacts, and feedback from practitioners and educators. Findings indicate that CoreTrust supports a shift from vague notions of community “needs” to spatially grounded reasoning about trade-offs across sustainability assets. The pilot also shows how practitioner feedback shapes design iterations and prepares the platform’s deliberative components for future citizen use.

Introduction

In recent decades, governments have redefined their relationships with citizens and moved towards smart governance by employing new digital technologies (Vigoda-Gadot & Mizrahi, 2024). Scholars highlight that accessible data is essential for stakeholder engagement and public trust (Mirzapour et al., 2023), yet city microdata often remains difficult to access and of uneven quality (Howe et al., 2022). Limited attention has been paid to how small and mid-sized cities (SMCs) can adopt them or how local practitioners and educators learn to interpret complex indicators. These gaps point to the need for learning-oriented, data-driven tools that support shared understanding of community data among stakeholders. Drawing on Learning Engineering (LE), an evidence-based, human-centered approach that iteratively designs, tests, and improves learning-supporting solutions in authentic contexts (Baker et al., 2022; Goodell et al., 2023), this paper examines CoreTrust as a data-driven civic learning tool for inclusive digital governance. We focus on CoreTrust, a platform grounded in the Community Capital Compass, a current articulation of the community capital framework in *Toward Sustainable Communities* (Roseland, Stout, & Spiliotopoulou, 2024). CoreTrust combines dashboards, digital twins, and deliberative components to support shared reasoning about community trade-offs and can be iteratively refined based on usage data and user feedback. To this end, this research aims to answer the following questions: (1) How can CoreTrust, as a data-driven civic learning tool, help educators and local practitioners interpret complex community data and understand sustainability trade-offs? (2) How can practitioner feedback from CoreTrust pilots inform iterative design improvements and support future citizen participation?

Methods

This study employs a qualitative design informed by participatory action research, tracing the iterative design and early implementation of the Community Capital Compass within CoreTrust and examines how this process functions as a civic learning environment for educators and local practitioners. To make the Learning Engineering connection explicit, we treat the workshop as a nested design–enactment–feedback cycle informing revisions (Totino & Kessler, 2024). The empirical focus is a pilot deployment of CoreTrust in Casa Grande, AZ, a SMC. The project is part of the Arizona Research Center for Housing and Economic Solutions (ARCHES), funded by the US Department of Housing and Urban Development (HUD). The pilot is developed through a collaboration between Arizona State University's Community Capital Lab (CCL) and the City of Casa Grande, with university-based educators and local practitioners engaging in workshop settings. Data for this study draw on three sources: (1) planning and policy documents, like city's general plan; (2) public data such as census data, embedded in the dashboard and digital twin with baseline measures of social, economic, and environmental assets at the block group level; and (3) design and feedback artifacts, including successive prototypes, internal design notes, and structured feedback from local practitioners on visualization of differences and trade-offs. Figure 1 illustrates the CoreTrust dashboard and 3D twin used in the Casa Grande pilot. Analysis proceeded through a qualitative, thematic review of project documentation, platform artifacts, and practitioner feedback to understand how educators and local practitioners used CoreTrust to interpret complex community data and trade-offs, and how actionable feedback and early usage patterns informed design changes to the dashboard, digital twin, and deliberative features.

Figure 1.

CoreTrust for the City of Casa Grande - Community Capital Overview dashboard (left). CG Downtown Twin used for 3D exploration of scenarios (right).



Results

CoreTrust helped planning commissioners move from abstract notions of community capital to a spatial understanding of city status. In scenario work, commissioners used the mapped indicators to identify where specific capitals were weak or strong, rather than discussing needs in the abstract. Participants compared the three sustainability assets across areas to articulate trade-offs while discussing development scenarios. In the CCL, students also worked with the pilot as part of their civic data learning. Moreover, Feedback from local practitioners led to design changes in CoreTrust. A key change was shifting data to the block-group level after planning commission members indicated that finer-grained geographies were more meaningful.

This change enabled more actionable, neighborhood-scale diagnosis during scenario discussions. Guided by the Community Capital Compass, the team first implemented affordable housing as an illustrative use case and then expanded the dashboard to additional assets, encouraging a multi-capital view of community conditions. Practitioners also requested street labels and clearer controls for indicator layers. For the deliberative technology, the citizen-facing layer, they noted that asking for precise addresses would discourage participation and recommended using ZIP code instead, informing the design of this component for future citizen participation.

Discussion

The findings from this pilot suggest that CoreTrust is beginning to function as a civic learning environment for both local practitioners and educators in a SMC context. For planning commission members, organizing existing indicators into community capital categories and visualizing them on a dashboard made the distribution of strengths and vulnerabilities across Casa Grande more legible. They moved from a general sense of “needs” to a more precise understanding of where each form of capital was weak or strong, and used the language of community capital when discussing trade-offs across social, economic, and environmental assets. From a learning-oriented design perspective, the pilot illustrates how feedback from local practitioners can guide iterative refinements to a digital data-driven tool without changing its underlying framework. Learning objectives vary by user group: supporting professional planning sensemaking for practitioners, civic data literacy for educators, and a future pathway for resident participation through the deliberative layer. These refinements surface design principles for scalable and inclusive civic tools: framing indicators through the Community Capital Compass, supporting users’ learning processes through clear language and guided interactions, and translating practice-based feedback into structured capitals that can be reused and compared over time. These results address the gaps discussed in the introduction. They demonstrate how a scalable and inclusive tool like CoreTrust can help SMCs move beyond static open-data portals toward learning-oriented, data-driven platforms that support shared sense-making. At the same time, this work remains at an early stage, focused on a single case and pre-citizen deployment. Future research will need to examine how residents use CoreTrust.

References

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it’s going, and the research needed. *Technology, Mind, and Behavior*. <https://doi.org/10.1037/tmb0000058>
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance: Based on the iFEST poster (winner of best poster design). *Journal of Military Learning*, 7(1), 46–59.
- Howe, B., Brown, J. M., Han, B., Herman, B., Weber, N., Yan, A., Yang, S., & Yang, Y. (2022). Integrative urban AI to expand coverage, access, and equity of urban data. *The European Physical Journal Special Topics*, 231(9–10), 1741–1752. <https://doi.org/10.1140/epjs/s11734-022-00475-z>
- Mirzapour, M., Tootian Isfahani, S., Memarzadeh, G. R., & Hashemzadeh Khorasgani, G. R. (2023). Identifying the constituent factors of open governance in public institutions. *International Journal of Human Capital in Urban Management*, 8(4), 545–558. <https://doi.org/10.22034/IJHCUM.2023.04.08>
- Roseland, M., Stout, M., Spiliotopoulou, M. (2024). *Toward Sustainable Communities: Solutions for Citizens and Their Governments*. Canada: New Society Publishers.
- Totino, L., & Kessler, A. (2024). “Why did we do that?” A Systematic Approach to Tracking Decisions in the Design and Iteration of Learning Experiences. *The Journal of Applied Instructional Design*, 13(2) <https://doi.org/10.59668/1269.15630>

Vigoda-Gadot, E., & Mizrahi, S. (2024). The digital governance puzzle: Towards integrative theory of humans, machines, and organizations in public management. *Technology in Society*, 77, 102530.
<https://doi.org/10.1016/j.techsoc.2024.102530>



Engineering a Semantic Topicality Instrument for Multiple-Choice Question Quality Control

Michelle Banawan, Linh Huynh, Katerina Christhlf, & Danielle S. McNamara

Distractor plausibility

Semantic similarity

Topic relevance

Ensuring that multiple-choice question (MCQ) distractors are conceptually relevant yet incorrect is critical for test validity and learner modeling. This study evaluates semantic topicality scoring as an instrument for automated MCQ quality control. A corpus of 20 reading-comprehension items (80 options) was annotated for topical relevance and analyzed using lexical and semantic similarity measures relative to passage main ideas. SBERT achieved the strongest correspondence with human judgments (Spearman's $\rho = 0.54$; AUC = 0.90). An empirically derived midpoint threshold classified 45% of distractors as on-topic, providing an interpretable operationalization of the rubric. Following a nested learning engineering process, we iteratively designed, implemented, and validated this semantic topicality instrument as an actionable diagnostic for AI-assisted item development. Results demonstrate that embedding-based similarity can serve as a scalable, human-aligned signal for content validity, supporting automated item screening, hybrid human–AI review, and future adaptive question-generation workflows.

Introduction

Multiple-choice questions (MCQs) remain among the most prevalent tools in educational assessment due to their efficiency, objectivity, ease of scoring, and reliability in large-scale testing contexts (Haladyna et al., 2002). Within the context of reading comprehension, high-quality MCQs are characterized by content relevance, clarity, and well-designed distractors that differentiate between varying levels of comprehension (Haladyna et al., 2002; Gorgun & Bulut, 2024). Effective items align with intended learning objectives, measure meaningful comprehension rather than surface recall, and employ plausible yet clearly incorrect distractors to maintain discriminative power.

Producing and reviewing such distractors is resource-intensive and difficult to scale. Item writers must ensure conceptual alignment with passage ideas, and quality reviewers must verify this alignment, both of which require time and expertise. As assessment programs increasingly rely on large item pools and AI-assisted item generation, there is a growing need for automated, interpretable methods to evaluate distractor topicality and support sustainable quality-control processes. Prior work shows that AI-generated MCQs often exhibit factual errors, reduced conceptual relevance, and limited higher-order reasoning (Cheung et al., 2023; Law et al., 2025), reinforcing the need for systematic topicality diagnostics regardless of item

source. Effective automation must therefore align with psychometric quality criteria such as appropriateness, clarity, relevance, discriminative power, and cognitive challenge (Feng et al., 2024).

To address this challenge, this study follows a Learning Engineering approach in which distractor topicality is treated as an engineering problem addressed through iterative cycles of design, instrumentation, implementation, and empirical investigation (Goodell, 2023; Craig et al., 2025). Consistent with Learning Engineering efforts to develop measurement frameworks within complex learning environments, we construct and validate a semantic topicality instrument using nested cycles of development and evaluation.

Methods

Dataset and Annotation

The dataset comprises 20 reading-comprehension questions evenly distributed between two subject domains: History and Lifespan Development. Each question includes one correct answer and three distractors, yielding 80 options (20 correct and 60 distractors). Each question is associated with a short passage and a main-idea summary written by reading comprehension experts to capture the essential conceptual content. These summaries serve as the reference text for measuring topical overlap between an option and its passage.

Options averaged 5 words and main-idea summaries averaged 82 words. Two expert annotators independently rated each option–main idea pair for topical relevance using a four-point ordinal rubric (Table 1).

Table 1.
Annotation Rubric for Topical Relevance

Rating	Description	Conceptual coverage
1	Disagree	Off-topic, does not address key concepts
2	Neutral	Peripheral, partially covered in sub-points
3	Agree	Addressed in sub-points of minor ideas
4	Strongly Agree	Fully covered in a main idea

Inter-rater agreement between the two annotators was 75% exact match (weighted Cohen’s $\kappa = 0.44$), indicating fair consistency in ordinal judgments of topical relevance. Disagreements occurred on five of the twenty items, typically involving adjacent category ratings (e.g., 3 vs 4). A deterministic majority score was derived for subsequent analyses by taking the mode of the two ratings (lower value in ties). The distribution of majority topical-relevance scores confirms that most answer options, including correct answers by design, were judged as highly relevant to their source text. This is consistent with the construction of pedagogically plausible distractors.

Similarity Metrics

To estimate topical relevance computationally, three complementary similarity measures were applied to each option–main idea pair. First, TF–IDF cosine similarity quantified the weighted lexical overlap between the two texts, emphasizing shared content words while down-weighting frequent or uninformative terms. Second, Jaccard overlap measured the ratio of shared

tokens to the total unique tokens across both texts, offering a simple indicator of vocabulary intersection. Finally, Sentence-BERT (SBERT) cosine similarity computed the cosine similarity between contextual embeddings generated by the all-MiniLM-L6-v2 model. This embedding-based method captures semantic proximity beyond surface wording, allowing conceptually related expressions to be recognized even when they share few lexical items. Together, these metrics allowed a direct comparison between lexical similarity (TF-IDF, Jaccard) and semantic similarity (SBERT) in predicting human topicality judgments.

Although main-idea summaries were substantially longer than the option texts, this asymmetry reflects authentic item structure. Embedding-based measures such as SBERT are robust to such length differences, whereas lexical metrics (TF-IDF, Jaccard) can exhibit mild sensitivity. These nuances were recognized and considered in the interpretation of results.

Evaluation Procedures

The alignment between computational measures and human ratings was examined through both correlational and classification-based analyses. Pearson's r and Spearman's ρ were computed to assess, respectively, linear and rank-order associations between similarity scores and human topical-relevance ratings. Ratings 3–4 were treated as on-topic and 1–2 as off-topic. Performance was assessed using empirically derived midpoint thresholds, defined as the midpoint between the mean similarity scores of correct and distractor options within each metric. For SBERT, this procedure yielded a midpoint threshold of 0.384, which was subsequently used for semantic classification analyses.

Results

Corpus Characteristics

Tokenization and part-of-speech tagging revealed that answer options were noun and adjective-dominant, while main-idea texts exhibited richer syntactic diversity. Readability analysis using the Flesch–Kincaid Grade Level metric showed that options averaged Grade 12.2, typical of tertiary-level assessments, whereas main-idea summaries averaged Grade 14.9, indicating dense, multi-sentence conceptual language. Vocabulary analysis identified 133 unique tokens in options and 251 in main-idea summaries, with minimal direct word overlap (mean Jaccard = 0.034). These findings confirm that topical relevance in this dataset depends primarily on semantic rather than lexical similarity, supporting its suitability for testing embedding-based models.

Correlation with Human Judgments

All three computational measures showed statistically significant positive associations with the human topical-relevance ratings ($p < .001$). The lexical approaches (TF-IDF and Jaccard) demonstrated moderate correlations with expert judgments, with Pearson's r values of 0.48 and 0.50, respectively, and comparable rank-order associations (Spearman's $\rho = 0.55$ for both). These results indicate that word-level overlap captures part of the relationship between option content and the passage's main ideas.

However, the SBERT semantic similarity measure exhibited the strongest correspondence with human ratings, achieving a Spearman's ρ of 0.54 and an area under the ROC curve (AUC) of 0.90, the highest among all tested models. This pattern suggests that the embedding-based approach more effectively represents the conceptual alignment that human raters perceive, particularly when lexical overlap is limited or paraphrased.

Table 3.

Classification performance at empirical thresholds

Measure	Threshold midpoint	Precision	Recall	F_1	Threshold conservative	Precision	Recall	F_1
TF-IDF	0.072	1.00	0.41	0.58	0.06	1.00	0.49	0.65
Jaccard	0.050	1.00	0.45	0.62	0.05	1.00	0.51	0.68
SBERT	0.384	1.00	0.50	0.67	0.38	1.00	0.51	0.68

Table 4.
On-Topic Classification using SBERT

Category	Count	% of Distractors
On-topic	27	45%
Off-topic	33	55%

Applying the SBERT threshold = 0.384 categorized 45% of distractors (27/60) as on topic. These corresponded to human relevance scores ≥ 3 , validating the threshold as a practical operationalization of rubric-based topical coverage.

Example Cases

A qualitative examination of individual items further illustrates the distinction between high- and low-scoring distractors identified by the semantic model.

Distractors receiving high SBERT similarity scores (≥ 0.60) typically paraphrased or recontextualized key concepts from the passage while maintaining overall conceptual fidelity. For instance, the option “Language acquisition is entirely dependent on social interactions and caregiver reinforcement” (SBERT = 0.69) closely reflects the thematic content of the passage on language development, emphasizing the role of social interaction. Similarly, the distractor “King James II successfully implemented a Catholic monarchy” (SBERT = 0.67) retains the central historical idea of religious absolutism underlying the text, despite presenting it in an overstated or misleading form.

In contrast, low-scoring distractors (SBERT < 0.15) tended to introduce information that was either tangential or factually irrelevant to the passage’s conceptual focus. Examples include “Experience is not necessary at all” (SBERT = 0.07), which directly contradicts the discussion of experiential learning in the source text, and “It improved relations with Native Americans” (SBERT = 0.09), which diverges entirely from the historical themes addressed. These comparisons demonstrate that high semantic similarity corresponds to distractors that remain topically grounded yet incorrect, whereas low similarity identifies options that deviate substantially from the main conceptual domain of the passage.

Correlations between similarity scores and answer correctness were weak ($r = 0.06\text{--}0.27$), confirming that topical relevance is distinct from factual correctness. This distinction reinforces the rubric’s focus on conceptual relevance rather than accuracy per se.

Discussion

The findings confirm that semantic embeddings capture conceptual relationships underlying human topicality judgments more effectively than word-overlap metrics. SBERT's performance (Spearman's $\rho = 0.54$, AUC = 0.90) demonstrates that embedding-based similarity can operationalize topic relevance in a scalable and interpretable way. The empirically derived midpoint threshold (0.384) provides a practical rule for identifying distractors that are conceptually aligned with a passage's key ideas. Correlations between similarity scores and answer correctness were weak ($r = 0.06$ – 0.27), confirming that topical relevance is distinct from factual accuracy—a distinction central to the rubric's focus on conceptual alignment rather than truth value.

The qualitative cases further illustrate the distinction that high-similarity distractors paraphrased or reframed key ideas while remaining conceptually grounded, whereas low-similarity options introduced tangential or irrelevant content. This pattern reinforces that high semantic similarity captures meaningful topical fidelity rather than mere wording overlap. Viewing this work through a Learning Engineering lens clarifies its next steps and practical value. The SBERT threshold provides an initial instrument for automated item screening, but real-world deployment requires iterative cycles of refinement: (a) scale testing across diverse domains, (b) psychometric linking of topicality to item difficulty/discrimination, (c) teacher-in-the-loop evaluation to calibrate tolerances for “plausible” distractors, and (d) integration into AI-item generation pipelines as a hybrid human–AI filter. Following these cycles will move the instrument from a promising diagnostic to a production-ready quality-control tool that aligns NLP diagnostics with educational validity requirements.

Conclusion and Future Work

This study demonstrated that computational similarity measures can approximate expert judgments of topical relevance in multiple-choice questions. Among the evaluated approaches, SBERT semantic embeddings showed the highest correspondence with human annotations (Spearman's $\rho = 0.54$, AUC = 0.90), outperforming lexical baselines such as TF–IDF and Jaccard overlap. These findings indicate that embedding-based topicality scoring can serve as a quantitative, interpretable indicator of content validity, supporting the automated review and generation of assessment items.

Future work will extend this analysis across larger and more diverse domains to test generalizability. Further, we will explore hybrid models that integrate lexical and semantic features, and evaluate how topicality scores relate to psychometric properties such as item difficulty and discrimination. Embedding-based topicality scoring could also be integrated into AI-assisted item development systems to flag weak and off-topic distractors, guide generative models, and provide conceptual diagnostics of learner misunderstanding. Collectively, these directions aim to refine topicality scoring as a scalable, human-aligned tool for maintaining the validity and quality of educational assessments.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305N210041 and R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

- Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., ... & Co, M. T. H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong SAR, Singapore, Ireland, and the United Kingdom). *PloS one*, 18(8), e0290691.
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale* (pp. 115-132). <https://doi.org/10.59668/2109.21735>

- Feng, W., Lee, J., McNichols, H., Scarlatos, A., Smith, D., Woodhead, S., ... & Lan, A. (2024, June). Exploring automated distractor generation for math multiple-choice questions via large language models. In Findings of the Association for Computational Linguistics: NAACL 2024 (pp. 3067-3082).
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning Engineering at a Glance. *Journal of Military Learning*.
- Gorgun, G., & Bulut, O. (2024). Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. *Education and Information Technologies*, 29, 24111–24142. <https://doi.org/10.1007/s10639-024-12771-3>
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–334. https://doi.org/10.1207/S15324818AME1503_5
- Law, A. K., So, J., Lui, C. T., Choi, Y. F., Cheung, K. H., Kei-ching Hung, K., & Graham, C. A. (2025). AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Medical Education*, 25(1), 208.



Design and Development of Two Digital Solutions for Promoting Harmonious Relationships: Buddy Up and Squeez

Kimberly R. M. Osborne, Ashley R. McDonald, Mina C. Johnson-Glenberg, Fuzhe Xie, Anqi Peng, Richard A. Fabes, Carol Lynn Martin, Sabina Low, & Laura D. Hanish

Digitized Interventions

Harmonious Relationships

Nested Learning Engineering Processes

Extended Abstract

Introduction

The Sanford Harmony Institute (SHI) is a new Institute at Arizona State University and this poster presents two important Works In Progress. The team leading SHI has created multiple hands-on and digitized solutions aimed at relationship-building and connectedness for children and the adults who care for them. Both apps were designed using best practices in learning engineering including: iterative design processes, co-designing with users (teachers and parents), and mixed methods (quantitative and qualitative) assessments on the efficacy of the tools (Goodell & Kolodner, 2023). This report shares the innovations and insights derived from one nested cycle within a broader, iterative learning engineering process (challenge-creation-implementation-investigation) for each digital tool during the creation phases of the projects (Craig et al., 2025; Totino & Kessler, 2024).

Tool 1: Buddy Up

Buddy Up is a mediated peer-pairing intervention that supports children's positive relationships with diverse others in a classroom. A previous, hard copy intervention was informed by Intergroup Contact Theory (Allport, 1954), which specifies the benefits of exposure to heterogeneous others. Past trials indicated the intervention was successful in its aims (Hanish et al., 2021, 2023). During the challenge phase of this project, teachers reported that it was difficult to track child pairings and to ensure that pairs were novel. This new digitized version automatically allows teachers to create groupings (usually pairs of 2) and randomly assigns children to the groupings while tracking who has been with whom over the course of one year.

Design

SHI partnered with graphic designers, teachers, and developers to iteratively refine the app. A new backend allows teachers to see the pairs created over the year and to easily reshuffle pairs. As a nested cycle in the creation phase, this app was user-tested in Spring 2025 with five teachers who had prior experience with the hard-copy version of Buddy Up.

Innovations and Insights

Two insights were unearthed. First, the protection of student data must be the foremost priority for researchers and designers, as it is for schools. A nicknaming convention was adopted to protect student names, servers are kept in the U.S., and student names are deleted after one year. Second, teachers' time on the app needed to be minimized. Adding a link to automatically import student names and creating nicknames was effective.

Tool 2: SQUEEZ

Squeeze is a mobile app offering 21 short games and activities for parents to play with their young children aged 3-7 years old whenever life's moments could use some fun or distraction (e.g., doctor's offices, grocery lines, etc.). The app serves as a digitized resource tool providing tailored, easy-to-use activities that fit the moment, removing guesswork and encouraging playful learning in daily routines. The 21 games come in three categories based on the location (e.g., quiet space) and energy level of the child (e.g. calm, etc.) and aim to strengthen the caregiver-child relationship and help children develop self-regulation skills. It is unique in that it is delivered with a smartphone, but NOT intended to be played on a smartphone.

Design

SHI researchers, who are experts in child development and caregiver-child relationships, collaborated with game developers and programmers to develop the app from hard-copy prototypes and iteratively refine the app. A beta version of Squeeze was evaluated for cultural and developmental appropriateness by 29 caregivers in Spring 2025.

Innovations and Insights

The design process included updating the Squeeze mascot, and adding features such as Squizmos, which are simple tools that provide ways to enhance enjoyment of the activities. Each activity card consists of an activity description, available squizmos, "New ways to Play" extensions, and a communication prompt to encourage persistence with the challenging aspects of the activities. Findings from beta testing indicated that caregivers desired a way to identify games according to age-appropriateness and for the games and activities to be more effectively described by the three category names, thus age icons and new category names were created.

Discussion

Following integration with the Harmony Academy website at National University, SHI's dissemination partner in bringing innovations to scale, the Buddy Up app will be released to a small group of teachers to undergo another round of user testing. Further refinement will follow this limited release and testing before the app is made available to all users of the Harmony Academy website. As for Squeeze, final usability testing was conducted in early 2026 to examine how the app was integrated into daily life. Data are currently being analyzed and the findings will inform final refinement before the Squeeze app will be ready for wide-scale dissemination. These two tools are part of a larger suite of tools and apps that fulfill the SHI mandate to create content that will help children and those in their lives have more harmonious interactions and relationships.

Acknowledgement

We would like to thank our partners, Codagami and Little 10 Ten Robots, for designing the apps.

References

Allport, G. W. (1954). The nature of prejudice. Reading, MA: Addison-Wesley.

Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). <https://doi.org/10.59668/2109.21735>

Goodell, J. & Kolodner, J. (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Routledge. <https://library.oapen.org/handle/20.500.12657/86250>

Hanish, L. D., Martin, C. L., Cook, R., DeLay, D., Lecheile, B., Fabes, R. A., Goble, P., & Bryce, C. (2021). Building integrated peer relationships in preschool classrooms: The potential of buddies. Journal of Applied Developmental Psychology, 73, 101257. <https://doi.org/10.1016/j.appdev.2021.101257>

Hanish, L. D., Xiao, S. X., Malouf, L. M., Martin, C. L., Goble, P., Fabes, R. A., DeLay, D., & Bryce, C. (2023). The Benefits of Buddies: Strategically Pairing Preschoolers with Other-Gender Classmates Promotes Positive Peer Interactions. Early Education and Development, 34(5), 1011–1025. <https://doi.org/10.1080/10409289.2022.2090773>

Totino, L., & Kessler, A. (2024). “Why did we do that?” A Systematic Approach to Tracking Decisions in the Design and Iteration of Learning Experiences. The Journal of Applied Instructional Design, 13(2) <https://doi.org/10.59668/1269.15630>



From Immersion to Action: How VR Influences Behavioral Intentions Around Advanced Water Purification

Ketevan Chachkhiani, Emily Machniak, Amber Hedquist, Poorva Ketkar, & Claire Lauer

Environmental Communication

Immersive Learning

Virtual Reality

This study explores how immersive Virtual Reality (VR) can shape public understanding of complex scientific processes, using Advanced Purified Water (APW) as a case example. Using a quasi-experimental pre-test/post-test design with independent samples, the study measures how participants' behavioral intentions change after they experience an educational VR simulation of the water purification process. Statistical tests (Chi-square, Fisher's Exact, and Mann–Whitney U) compare pre- and post-test responses. Results show that the VR experience increased participants' willingness and comfort with using APW, demonstrating how VR can build understanding and trust in emerging technologies that the public will increasingly rely on.

From Immersion to Action: How VR Influences Behavioral Intentions Around Advanced Water Purification

Ketevan Chachkhiani¹, Emily Machniak², Amber Hedquist³, Poorva Ketkar⁴, Claire Lauer⁵

1 Arizona State University; kchachkh@asu.edu

2 Arizona State University; emachnia@asu.edu
anhedqui@asu.edu

3 Arizona State University;

4 Arizona State University; pketkar@asu.edu

5 Arizona State University; claire.lauer@asu.edu

Abstract.

Keywords: Virtual Reality, Immersive Learning, Environmental Communication

Introduction

Virtual reality (VR) is rapidly emerging as a transformative tool for science and environmental communication because of its ability to generate deep psychological engagement and behavioral change (Behm-Morawitz & Shin, 2024; Fauville et al., 2020; Breves, 2020). Drawing from the Theory of Planned Behavior (Ajzen, 1985), which posits that attitudes and perceived control predict intentions and subsequent actions, VR uniquely influences both cognitive and affective components of behavior. Immersion creates a powerful sense of presence, enabling audiences to experience scientific phenomena rather than merely observe them (Behm-Morawitz & Shin, 2024; Breves, 2020). This presence enhances message credibility, emotional connection, and empathy, producing persuasive effects (Breves & Schramm, 2021; Ahn et al., 2014). These mechanisms allow VR to bridge the gap between abstract information and lived experience, engaging both emotional and rational pathways that underlie behavioral intention (Behm-Morawitz & Shin, 2024; Ahn et al., 2014; Breves, 2020; Markowitz & Bailenson, 2021; Fauville et al., 2020).

A growing body of research demonstrates that VR-based experiences can enhance environmental learning, emotional engagement, and pro-environmental behavioral intentions. Studies show that immersive simulations reduce psychological distance and help participants visualize the tangible impacts of environmental issues, thereby strengthening concern and action tendencies (Markowitz & Bailenson, 2021; Luzzati et al., 2025). For example, Luzzati et al. (2025) found that participants who explored a climate-change scenario in VR reported greater perceived urgency and stronger intentions to adopt sustainable behaviors than those receiving text-based information. Similarly, Ahn et al. (2014) showed that embodiment in virtual environments increased participants' environmental locus of control, translating awareness into behavioral intentions. Experiments using 360° and embodied simulations report comparable effects: Fauville et al. (2020) demonstrated that participants who experienced ocean acidification through VR exhibited higher environmental concern than control groups viewing the same content in 2D, and Shin and Lee (2024) found that immersive campaign videos prompted deeper cognitive elaboration and alignment with pro-environmental messaging. Meta-analytic studies also substantiate these findings. Villena-Taranilla et al. (2022) synthesized 54 studies and identified a large effect of VR on motivation and learning outcomes, while Zhang and Bowman (2022) confirmed that interactive storytelling in VR enhanced both engagement and content retention. Overall, these studies establish VR as a powerful experiential medium that can influence cognition, emotion, and behavior in sustainability contexts. Furthermore, ethical design frameworks further emphasize that persuasive effectiveness depends on user agency, narrative coherence, and representational transparency (Lambrecht et al., 2024, 2025).

While VR's potential to change attitudes and behaviors is well established in environmental and educational contexts, little research has explored how it might help the public understand and accept specific sustainability technologies they are being asked to adopt. To address this gap, the current study examines VR's persuasive impact on the environmental science topic of APW production. Although APW systems have been proven safe and environmentally beneficial, messaging from municipalities has failed to account for residents' cognitive and affective barriers and has thus failed to convince a skeptical public that returning recycled wastewater directly back to the public drinking system is safe (Annin, 2024; Dolnicar et al., 2011). Phrases like, "toilet to tap," for instance, have historically hindered AWP development (Annin, 2024). This makes AWP a high-risk sustainability issue because despite it being a "viable near-term solution" for water security in Arizona, perceptual barriers

remain and could reduce public acceptance. The psychological and perceptual hurdles that prevent people from feeling comfortable adopting a new sustainability technology like Advanced Purified Water is just the sort of visceral challenge that the embodied nature of VR may help overcome. By bringing people directly into the water-purification processes to demonstrate its safety, VR can reduce perceived risk, foster understanding, and strengthen acceptance. Therefore, this study asks whether participants' use of VR to explore an APW system produces changes in behavioral intention, which, according to the Theory of Planned Behavior (Ajzen, 1991), is operationalized in this study through self-reported comfort and willingness. These measures function as theoretically informed indicators of attitudinal evaluation and perceived behavioral control shaping intention, consistent with established TPB measurement practices.

Methods

Research Design

Guided by a Learning Engineering framework, this study employs a human-centered, evidence-driven approach to evaluate an immersive VR learning intervention (Goodell et al., 2023). The research design represents a focused evaluation cycle embedded within a broader design effort, consistent with nested Learning Engineering methodologies (Craig et al., 2025). The study used a quasi-experimental pre-test/post-test design with independent samples to examine the impact of a VR educational experience on participants' willingness and comfort using APW. Participants played Flow Forward, a 15-minute VR game created by members of the Arizona Water Innovation Initiative and students in ASU's MESH Lab ("ASU Mesh Labs," 2025). In the game, players learn about advanced water purification, then pilot a nanosubmarine through ozonation, ultrafiltration, reverse osmosis, and UV filtration to remove contaminants and complete the process. Participants completed a brief online survey before (pre-test) and after (post-test) engaging in a VR experience illustrating the Advanced Water Purification process. Because individual pre- and post-responses could not be linked, comparisons were made between two independent groups.

Data Collection and Analysis

Researchers obtained IRB approval to work with high school students (IRB STUDY #00018372). Participants were recruited through a convenience sample consisting of two sections of AP Environmental Science during the spring semester of 2025. This population was particularly relevant because students were concurrently learning about water systems and environmental sustainability, making Advanced Water Purification a familiar and meaningful topic of study. After entering the classrooms, the project lead explained the study and administered the online pre-survey. The pre-survey defined APW as a process "to recycle and treat wastewater so that it is clean enough to drink" and asked students to specify, on a four-point scale from extremely uncomfortable to extremely comfortable, the extent to which they would feel comfortable drinking APW water or using it for a variety of other purposes (see Table 2). The development lead then provided a short walkthrough of the headset and general flow of the game before students used MetaQuest headsets to play the game from start to finish. When done, students filled out the post-survey where they were, once again, asked to what extent they would feel comfortable drinking APW and using it for other domestic purposes, before taking part in brief, informal group debriefs about their experiences.

Data analysis followed a structured process to address the three research questions. For RQ1, a Chi-square test of independence and Fisher's Exact Test were conducted to assess differences in willingness to drink APW between pre- and post-test groups, with Cramer's V reported as an effect size. For RQ2, independent-samples Mann-Whitney U tests were conducted to compare comfort ratings across the five APW uses, with effect sizes (r) calculated as $Z/\sqrt{N}$. For RQ3, effect sizes across uses were compared to determine which APW use showed the largest change in comfort. All analyses emphasized effect sizes and trends rather than significance alone, given the small sample size.

Results

Willingness to Drink Advanced Purified Water (RQ 1)

Out of 32 students, two students opted not to fully participate. All other students completed the pre-survey, gameplay, and post-survey. A cross-tabulation examined whether exposure to the VR experience influenced participants' willingness to drink Advanced Purified Water (APW). In the pre-test group, 26 of 30 participants (86.7%) reported 'Yes' while four (13.3%) said 'Maybe'. In the post-test group, all 30 participants (100%) reported 'Yes'. A Chi-square test of independence indicated a statistically significant association between group (pre vs. post) and willingness to drink APW, $\chi^2(1, N = 60) = 4.29, p = .038$, with a moderate effect size, Cramer's $V = .27$. Because 50% of the cells had expected counts less than five, Fisher's Exact Test was examined as a more conservative test. Fisher's test was not significant on the two-sided comparison ($p = .112$) but approached significance on the one-sided test ($p = .056$). These results suggest a positive trend toward greater willingness to drink APW following the VR experience (see Table 1).

Table 1.

Willingness to Drink Advanced Purified Water (Pre vs. Post VR)

Group	Yes	Maybe	No	Total (n)
Pre-test	26 (86.7%)	4 (13.3%)	0 (0%)	30
Post-test	30 (100%)	0 (0%)	0 (0%)	30
Total	56 (93.3%)	4 (6.7%)	0 (0%)	60

Statistical Test: $\chi^2(1, N = 60) = 4.29, p = .038$; Fisher's Exact (two-sided) = .112, one-sided = .056
Effect Size: Cramer's $V = .27$ (moderate)

Comfort Using Advanced Purified Water (RQ 2)

Descriptive statistics and Mann–Whitney U tests were conducted to examine whether comfort using APW increased after the VR experience across five household purposes. Median comfort ratings were high for both groups, generally ranging between 3 and 4 on a 4-point scale. A significant difference was found for drinking, $U = 586.50, Z = 2.19, p = .028, r = .28$, indicating greater comfort after the VR experience. No significant differences were found for cooking ($p = .422$), showering ($p = .137$), laundry ($p = .679$), or gardening ($p = .920$). The moderate effect for drinking suggests that the VR experience enhanced comfort primarily for the most personal use of APW (See Table 2).

Table 2.

Comfort with Using Advanced Purified Water Across Purposes (Pre vs. Post VR)

Purpose	Pre Median	Post Median	U	Z	p (2-tailed)	r
Drinking	3	4	586.50	2.19	.028	.28
Cooking	4	4	496.00	0.80	.422	.10

Showering	4	4	534.00	1.49	.137	.19
Laundry	4	4	473.00	0.41	.679	.05
Gardening	4	4	445.00	-0.10	.920	.01

Which APW Uses Show the Largest Change in Comfort? (RQ 3)

Effect size comparisons revealed that comfort for drinking showed the largest positive change following the VR experience ($r = .28$), followed by a smaller trend for showering ($r = .19$). Cooking, laundry, and gardening showed negligible changes ($r < .10$). This pattern indicates that the VR experience most strongly influenced comfort for direct personal uses such as drinking, supporting the potential of immersive educational tools to promote behavioral change toward water reuse.

Discussion

This study demonstrates how VR can make complex sustainability processes like APW production understandable and engaging. The findings suggest that comfort and willingness to incorporate APW in daily life tend to increase after the VR experience. In particular, there was a greater willingness to drink APW (RQ 1), with fewer participants being less uncertain about their willingness following the VR experience. Comfort using APW (RQ 2) increased, but only for drinking purposes (RQ 3), as using APW across other purposes was already largely accepted and thus did not change following the VR experience.

By combining immersion with narrative-driven interaction, the experience helped participants connect large-scale issues such as water scarcity and climate change to their own lives (Petersen et al., 2020; Lambrecht et al., 2025). Positioning users to “battle” contaminants through each purification stage also appeared to reduce psychological distance and foster active engagement, echoing findings from Markowitz and Bailenson (2021). Our study team will report in future work on additional variables, including participants’ most memorable moments and reflections from the VR experience. More broadly, future research can examine how narrative design and emotional engagement in VR foster lasting awareness and real-world action.

Limitations

This study provides early evidence of how immersive VR can shape public perceptions, but several limitations should be noted: (1) because individual pre- and post-test responses could not be linked, the analysis focused on group-level trends rather than participant-specific changes; (2) the sample consisted solely of high school students from similar educational backgrounds, limiting the generalizability of findings to broader populations.

Conclusion

Findings from this study suggest that immersive, narrative-based VR experiences can positively influence behavioral intentions related to sustainable technologies. By situating this evaluation within a Learning Engineering framework, the study demonstrates how iterative, human-centered design and empirical testing can inform the development of effective immersive learning interventions for real-world sustainability challenges. Future research will extend this work by examining participants’ learning, engagement, and perceptions within the same VR environment.

Acknowledgments

The authors would like to thank ASU's MESH Lab for developing the Flow Forward VR experience. This work was supported by Arizona State University's Arizona Water Innovation Initiative and the Impact Water—Arizona program, courtesy of the State of Arizona and the Virginia G. Piper Charitable Trust. Additional support was provided by the National Science Foundation (Grant Number 2314259).

References

Arizona Department of Environmental Quality. (2025, July 8). Advanced water purification.

<https://www.azdeq.gov/water-quality-division/advanced-water-purification>

Ahn, S. J. G., Bailenson, J. N., & Park, D. (2014). "Short-and long-term effects of embodied experiences in immersive virtual environments on environmental locus of control and behavior." *Computers in Human Behavior*, 39, 235-245. Ajzen, Icek. "From intentions to actions: A theory of planned behavior." *Action control: From cognition to behavior*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1985.11-39. Beckmann (Eds.), *Action control: From cognition to behavior* (pp. 11–39). Springer.

Annin, P. (2023). *Purified: How Recycled Sewage Is Transforming Our Water*. Island Press.

ASU Mesh Labs students create award-winning VR experience to educate about water re-use. (2025). ASU News.

<https://news.asu.edu/b/20251031-asu-mesh-labs-students-create-awardwinning-vr-experience-educate-about-water-reuse>

Behm-Morawitz, E., & Shin, H. (2024). "Using VR for science communication: Presence, message perception, and pro-environmental effects." *Science Comm.*, 46(1), 36–64.

Breves, P. (2020). "Bringing people closer: The prosocial effects of immersive media on users' attitudes and behavior." *Nonprofit and Voluntary Sector Quarterly*, 49(5), 1015–1034.

Breves, P., & Schramm, H. (2021). "Bridging psychological distance: The impact of immersive media on distant and proximal environmental issues." *Computers in Human Behavior*, 115, 106606.

Craig, S. D., Avancha, K., Malhotra, P., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a nested learning engineering methodology to develop a team dynamic measurement framework for a virtual training environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings* (pp. 115–132).

Dolnicar, S., Hurlimann, A., & Grün, B. (2011). "What affects public acceptance of recycled and desalinated water?" *Water Research*, 45(2), 933–943.

Fauville, G., Queiroz, A. C. M., & Bailenson, J. N. (2020). "Virtual reality as a promising tool to promote climate change awareness." In J. Jim & H. Song (Eds.), *Technology and health: Promoting attitude and behavior change* (pp. 91–108).

Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. *Journal of Military Learning*. <https://www.armyupress.army.mil/Journals/Journal-of-Military-Learning/Journal-of-Military-Learning-Archives/Conference-Edition-2023-Journal-of-Military-Learning/Engineering-at-a-Glance/>

Lambrecht, K., Carradini, S., Lauer, C., & Mara, A. (2024, October). "Developing an ethical framework for virtual reality design and implementation in technical communication." In *Proceedings of the 42nd ACM International Conference on Design*

of Communication (pp. 37–44). Association for Computing Machinery. <https://doi.org/10.1145/3674327.3674333>

Lambrech, K., Lauer, C., Carradini, S., & Ketkar, P. (2025). "Telling Technical Communication Stories through Virtual Reality." *Proceedings of the 43rd ACM International Conference on Design of Communication*, 114–120. <https://doi.org/10.1145/3711670.3764627>

Luzzati, T., Baraldi, S., Ermini, S., Fata, C., Faralla, V., Guarnieri, P., Lusuadi, L., Santalucia, V., Scipioni, S., Sirizzotti, M., & Innocenti, A. (2025). "Can improving climate change perception lead to more environmentally friendly choices? Evidence from an immersive virtual environment experiment." *Ecological Economics*, 229.

Markowitz, D. M., & Bailenson, J. N. (2021). "Virtual reality and the psychology of climate change." *Current Opinion in Psychology*, 42, 60–65.

Petersen, G. B., Klingenberg, S., Mayer, R. E., & Makransky, G. (2020). "The virtual field trip: Investigating how to optimize immersive virtual learning in climate change education." *British Journal of Educational Technology*, 51(6), 2099–2115. <https://doi.org/10.1111/bjet.12991>

Shin, M., & Lee, H. (2024). "Harnessing 360-degree video to prompt users to think along with pro-environmental campaign messages." *Media and Comm.*, 12(2), 1–10.

Villena-Taranilla, R., Moreno-Guerrero, A. J., & Fernández-Cerero, J. (2022). "Virtual reality as a resource for improving learning outcomes and motivation in STEM education: A meta-analysis." *Education and Information Technologies*, 27(5), 6773–6799.

Zhang, L., & Bowman, D. A. (2022). "Exploring effect of level of storytelling richness on science learning in immersive virtual reality." *Proceedings of the 2022 ACM International Conference on Interactive Media Experiences*, 19–32.



From Immersion to Behavioral Intentions: Virtual Reality and Advanced Water Purification

Ketevan Chachkhiani, Emily Machniak, Amber Hedquist, Poorva Ketkar, & Claire Lauer

Environmental Communication

Immersive Learning

Virtual Reality

This study examines the relationship between immersive Virtual Reality (VR) and behavioral intentions related to Advanced Purified Water (APW). Using a quasi-experimental pre-test/post-test design, the study compares participants' willingness and comfort using APW before and after engagement with an educational VR simulation of the water purification process. Statistical tests (Chi-square, Fisher's Exact, and Mann–Whitney U) were used to compare pre- and post-test responses. Results show that overall participants' willingness and comfort with using APW were higher following the VR experience, suggesting an association between immersive educational VR and increased acceptance of emerging water technologies.

Introduction

Virtual reality (VR) is rapidly emerging as a transformative tool for science and environmental communication because of its ability to generate deep psychological engagement and behavioral change (Behm-Morawitz & Shin, 2024; Fauville et al., 2020; Breves, 2020). Drawing from the Theory of Planned Behavior (Ajzen, 1985), which posits that attitudes and perceived control predict intentions and subsequent actions, VR uniquely influences both cognitive and affective components of behavior. Immersion creates a powerful sense of presence, enabling audiences to experience scientific phenomena rather than merely observe them (Behm-Morawitz & Shin, 2024; Breves, 2020). This presence enhances message credibility, emotional connection, and empathy, producing persuasive effects (Breves & Schramm, 2021; Ahn et al., 2014). These mechanisms allow VR to bridge the gap between abstract information and lived experience, engaging both emotional and rational pathways that underlie behavioral intention (Behm-Morawitz & Shin, 2024; Ahn et al., 2014; Breves, 2020; Markowitz & Bailenson, 2021; Fauville et al., 2020).

A growing body of research demonstrates that VR-based experiences can enhance environmental learning, emotional engagement, and pro-environmental behavioral intentions. Studies show that immersive simulations reduce psychological distance and help participants visualize the tangible impacts of environmental issues, thereby strengthening concern and action tendencies (Markowitz & Bailenson, 2021; Luzzati et al., 2025). For example, Luzzati et al. (2025) found that participants who explored a climate-change scenario in VR reported greater perceived urgency and stronger intentions to adopt sustainable behaviors than those receiving text-based information. Similarly, Ahn et al. (2014) showed that embodiment in virtual environments increased participants' environmental locus of control, translating awareness into behavioral intentions. Experiments using 360° and embodied simulations report comparable effects: Fauville et al. (2020) demonstrated that participants who experienced ocean acidification through VR exhibited higher environmental concern than control groups viewing the same content in 2D, and Shin and Lee (2024) found that immersive campaign videos prompted deeper cognitive elaboration and alignment with pro-environmental messaging. Meta-analytic studies also substantiate these findings. Villena-Taranilla et al. (2022) synthesized 54 studies and identified a large effect of VR on motivation and learning outcomes, while Zhang and Bowman (2022) confirmed that interactive storytelling in VR enhanced both engagement and content retention. Overall, these studies establish VR as a powerful experiential medium that can influence cognition, emotion, and behavior in sustainability contexts. Furthermore, ethical design frameworks further emphasize that persuasive effectiveness depends on user agency, narrative coherence, and representational transparency (Lambrecht et al., 2024, 2025).

While VR's potential to change attitudes and behaviors is well established in environmental and educational contexts, little research has explored how it might help the public understand and accept specific sustainability technologies they are being asked to adopt. To address this gap, the current study examines how immersive VR relates to public acceptance and behavioral intentions toward AWP. Although APW systems have been proven safe and environmentally beneficial, messaging from municipalities has failed to account for residents' cognitive and affective barriers and has thus failed to convince a skeptical public that returning recycled wastewater directly back to the public drinking system is safe (Annin, 2024; Dolnicar et al., 2011). Phrases like, "toilet to tap," for instance, have historically hindered AWP development (Annin, 2024). This makes AWP a high-risk sustainability issue because despite it being a "viable near-term solution" for water security in Arizona, perceptual barriers remain and could reduce public acceptance. The psychological and perceptual hurdles that prevent people from feeling comfortable adopting a new sustainability technology like Advanced Purified Water is just the sort of visceral challenge that the embodied nature of VR may help overcome. By bringing people directly into the water-purification processes to demonstrate its safety, VR can reduce perceived risk, foster understanding, and strengthen acceptance. Therefore, this study asks whether participants' engagement with a VR exploration of an APW system is associated with differences in behavioral intentions, which, according to the Theory of Planned Behavior (Ajzen, 1985), is operationalized in this study through self-reported comfort and willingness. These measures function as theoretically informed indicators of attitudinal evaluation and perceived behavioral control shaping intention, consistent with established TPB measurement practices.

Methods

Research Design

Guided by a Learning Engineering framework, this study employs a human-centered, evidence-driven approach to evaluate an immersive VR learning intervention (Goodell et al., 2023). The research design represents a focused evaluation cycle embedded within a broader design effort, consistent with nested Learning Engineering methodologies (Craig et al., 2025). The study used a quasi-experimental pre-test/post-test design to examine differences in participants' willingness and comfort using APW between pre- and post-VR groups. Participants played Flow Forward, a 15-minute VR game created by members of the Arizona Water Innovation Initiative and students in ASU's MESH Lab ("ASU Mesh Labs," 2025). In the game, players learn about advanced water purification, then pilot a nanosubmarine through ozonation, ultrafiltration, reverse osmosis, and UV filtration to remove contaminants and complete the process. Participants completed a brief online survey before (pre-test) and after (post-test) engaging in a VR experience illustrating the Advanced Water Purification process. Since individual participants' pre- and post-test responses could not be matched, pre- and post-test scores were treated as independent samples for analysis.

Data Collection and Analysis

Researchers obtained IRB approval to work with high school students (IRB STUDY #00018372). Participants were recruited through a convenience sample consisting of two sections of AP Environmental Science during the spring semester of 2025. This population was particularly relevant because students were concurrently learning about water systems and environmental sustainability, making Advanced Water Purification a familiar and meaningful topic of study. After entering the classrooms, the project lead explained the study and administered the online pre-survey. The pre-survey defined APW as a process "to recycle and treat wastewater so that it is clean enough to drink" and asked students to specify, on a four-point scale from extremely uncomfortable to extremely comfortable, the extent to which they would feel comfortable drinking APW water or using it for a variety of other purposes (see Table 2). The development lead then provided a short walkthrough of the headset and general flow of the game before students used MetaQuest headsets to play the game from start to finish. When done, students filled out the post-survey where they were, once again, asked to what extent they would feel comfortable drinking APW and using it for other domestic purposes, before taking part in brief, informal group debriefs about their experiences. Out of 32 students, two students opted not to fully participate. All other students completed the pre-survey, gameplay, and post-survey.

Data analysis followed a structured process to address the three research questions. For RQ1, a Chi-square test of independence and Fisher's Exact Test were conducted to assess differences in willingness to drink APW between pre- and post-test groups, with Cramer's V reported as an effect size. For RQ2, independent-samples Mann-Whitney U tests were conducted to compare comfort ratings across the five APW uses, with effect sizes (r) calculated as $Z/\sqrt{N}$. For RQ3, effect sizes across uses were compared to determine which APW use showed the largest change in comfort. All analyses emphasized effect sizes and trends rather than significance alone, given the small sample size.

Results

Willingness to Drink Advanced Purified Water (RQ 1)

A cross-tabulation examined whether exposure to the VR experience exposure to an immersive VR experience was associated with greater willingness to drink APW. In the pre-test group, 26 of 30 participants (86.7%) reported 'Yes' while four (13.3%) said 'Maybe'. In the post-test group, all 30 participants (100%) reported 'Yes'. A Chi-square test of independence indicated a statistically significant association between group (pre vs. post) and willingness to drink APW, $\chi^2(1, N = 60) = 4.29, p = .038$, with a moderate effect size, Cramer's $V = .27$. Because 50% of the cells had expected counts less than five, Fisher's Exact Test was examined as a more conservative test. Fisher's test was not significant on the two-sided comparison ($p = .112$) but

approached significance on the one-sided test ($p = .056$). Overall, these results suggest a positive trend toward greater willingness to drink APW following the VR experience (see Table 1).

Table 1.

Willingness to Drink Advanced Purified Water (Pre vs. Post VR)

Group	Yes	Maybe	No	Total (n)
Pre-test	26 (86.7%)	4 (13.3%)	0 (0%)	30
Post-test	30 (100%)	0 (0%)	0 (0%)	30

Statistical Test: $\chi^2(1, N = 60) = 4.29, p = .038$; Fisher's Exact (two-sided) = .112, one-sided = .056 Effect Size: Cramer's $V = .27$

Comfort Using Advanced Purified Water (RQ 2)

Descriptive statistics and Mann–Whitney U tests were used to examine differences in comfort using APW between pre- and post-VR groups across five household purposes. Median comfort ratings were high for both groups, generally ranging between 3 and 4 on a 4-point scale. A significant difference was found for drinking, $U = 586.50, Z = 2.19, p = .028, r = .28$, indicating greater comfort after the VR experience. No significant differences were found for cooking ($p = .422$), showering ($p = .137$), laundry ($p = .679$), or gardening ($p = .920$). The moderate effect for drinking indicates greater comfort with the most personal use of APW in the post-VR group compared to the pre-VR group (see Table 2).

Table 2.

Comfort with Using Advanced Purified Water Across Purposes (Pre vs. Post VR)

Purpose	Pre Median	Post Median	U	Z	p (2-tailed)	r
Drinking	3	4	586.50	2.19	.028	.28
Cooking	4	4	496.00	0.80	.422	.10
Showering	4	4	534.00	1.49	.137	.19
Laundry	4	4	473.00	0.41	.679	.05
Gardening	4	4	445.00	-0.10	.920	.01

Change in Comfort After the VR Experience by APW Use (RQ 3)

Effect size comparisons showed that comfort for drinking exhibited the largest difference between pre- and post-VR groups ($r = .28$), followed by a smaller difference for showering ($r = .19$). Comfort for cooking, laundry, and gardening showed

negligible differences ($r < .10$). Overall, these patterns suggest that differences in comfort were greatest for direct, personal uses of APW, supporting the potential of immersive educational tools to promote behavioral change toward water reuse.

Discussion

This study demonstrates how VR can make complex sustainability processes like APW production understandable and engaging. The findings suggest that comfort and willingness to incorporate APW in daily life tend to increase after the VR experience. In particular, there was a greater willingness to drink APW (RQ 1), with fewer participants being less uncertain about their willingness following the VR experience. Comfort using APW (RQ 2) increased, but only for drinking purposes (RQ 3), as using APW across other purposes was already largely accepted and thus did not change following the VR experience.

By combining immersion with narrative-driven interaction, the experience helped participants connect large-scale issues such as water scarcity and climate change to their own lives (Petersen et al., 2020; Lambrecht et al., 2025). Positioning users to “battle” contaminants through each purification stage shows potential to reduce psychological distance and foster active engagement, echoing findings from Markowitz and Bailenson (2021). Future work will extend this research by systematically isolating and examining specific VR design features to better understand their individual and combined effects on participant outcomes. Additional variables, including emotional engagement, sense of presence, empathy, perceived agency, memorable moments, and participant reflections, can be incorporated to clarify the mechanisms through which immersive VR may shape behavioral intentions.

Limitations

This study provides early evidence of how immersive VR can shape public perceptions, but several limitations should be noted: (1) because individual pre- and post-test responses could not be linked, the analysis focused on group-level trends rather than participant-specific changes; (2) the sample consisted solely of high school students from similar educational backgrounds, limiting the generalizability of findings to broader populations.

Conclusion

Findings from this study suggest that immersive, narrative-based VR experiences have potential to positively influence behavioral intentions related to sustainable technologies. By situating this evaluation within a Learning Engineering framework, the study demonstrates how iterative, human-centered design and empirical testing can inform the development of effective immersive learning interventions for real-world sustainability challenges. Future research will extend this work by examining participants’ learning, engagement, and perceptions within the same VR environment.

Acknowledgments

The authors would like to thank ASU’s MESH Lab for developing the Flow Forward VR experience. This work was supported by Arizona State University’s Arizona Water Innovation Initiative and the Impact Water—Arizona program, courtesy of the State of Arizona and the Virginia G. Piper Charitable Trust. Additional support was provided by the National Science Foundation (Grant Number 2314259).

References

Arizona Department of Environmental Quality. (2025, July 8). Advanced water purification.

<https://www.azdeq.gov/water-quality-division/advanced-water-purification>

Ahn, S. J. G., Bailenson, J. N., & Park, D. (2014). "Short-and long-term effects of embodied

experiences in immersive virtual environments on environmental locus of control and behavior." *Computers in Human Behavior*, 39, 235-245.

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In *Action control: From cognition to behavior* (pp. 11-39). Berlin, Heidelberg: Springer Berlin Heidelberg.

Annin, P. (2023). *Purified: How Recycled Sewage Is Transforming Our Water*. Island Press.

ASU Mesh Labs students create award-winning VR experience to educate about water

re-use. (2025). ASU News. <https://news.asu.edu/b/20251031-asu-mesh-labs-students-create-awardwinning-vr-experience-educate-about-water-reuse>

Behm-Morawitz, E., & Shin, H. (2024). "Using VR for science communication: Presence, message perception, and pro-environmental effects." *Science Comm.*, 46(1), 36–64.

Breves, P. (2020). "Bringing people closer: The prosocial effects of immersive media on users' attitudes and behavior." *Nonprofit and Voluntary Sector Quarterly*, 49(5), 1015–1034.

Breves, P., & Schramm, H. (2021). "Bridging psychological distance: The impact of

immersive media on distant and proximal environmental issues." *Computers in Human Behavior*, 115, 106606. Craig, S. D., Avancha, K., Malhotra, P., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a nested learning engineering methodology to develop a team dynamic measurement framework for a virtual training environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings* (pp. 115–132).

Dolnicar, S., Hurlimann, A., & Grün, B. (2011). "What affects public acceptance of recycled and desalinated water?" *Water Research*, 45(2), 933–943.

Fauville, G., Queiroz, A. C. M., & Bailenson, J. N. (2020). "Virtual reality as a promising

tool to promote climate change awareness." In J. Jim & H. Song (Eds.), *Technology and health: Promoting attitude and behavior change* (pp. 91–108).

Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. *Journal of Military Learning*. <https://www.armyupress.army.mil/Journals/Journal-of-Military-Learning/Journal-of-Military-Learning-Archives/Conference-Edition-2023-Journal-of-Military-Learning/Engineering-at-a-Glance/>

Lambrecht, K., Carradini, S., Lauer, C., & Mara, A. (2024, October). "Developing an

ethical framework for virtual reality design and implementation in technical communication." In *Proceedings of the 42nd ACM International Conference on Design of Communication* (pp. 37–44). Association for Computing Machinery. <https://doi.org/10.1145/3674327.3674333>

- Lambrecht, K., Lauer, C., Carradini, S., & Ketkar, P. (2025). "Telling Technical Communication Stories through Virtual Reality." *Proceedings of the 43rd ACM International Conference on Design of Communication*, 114–120. <https://doi.org/10.1145/3711670.3764627>
- Luzzati, T., Baraldi, S., Ermini, S., Faita, C., Faralla, V., Guarnieri, P., Lusuardi, L., Santalucia, V., Scipioni, S., Sirizzotti, M., & Innocenti, A. (2025). "Can improving climate change perception lead to more environmentally friendly choices? Evidence from an immersive virtual environment experiment." *Ecological Economics*, 229.
- Markowitz, D. M., & Bailenson, J. N. (2021). "Virtual reality and the psychology of climate change." *Current Opinion in Psychology*, 42, 60–65.
- Petersen, G. B., Klingenberg, S., Mayer, R. E., & Makransky, G. (2020). "The virtual field trip: Investigating how to optimize immersive virtual learning in climate change education." *British Journal of Educational Technology*, 51(6), 2099–2115. <https://doi.org/10.1111/bjet.12991>
- Shin, M., & Lee, H. (2024). "Harnessing 360-degree video to prompt users to think along with pro-environmental campaign messages." *Media and Comm.*, 12(2), 1–10.
- Villena-Taranilla, R., Moreno-Guerrero, A. J., & Fernández-Cerero, J. (2022). "Virtual reality as a resource for improving learning outcomes and motivation in STEM education: A meta-analysis." *Education and Information Technologies*, 27(5), 6773–6799.
- Zhang, L., & Bowman, D. A. (2022). "Exploring effect of level of storytelling richness on science learning in immersive virtual reality." *Proceedings of the 2022 ACM International Conference on Interactive Media Experiences*, 19–32.



How Students Really Use Courseware: Visualizing Student Pathways in Integrated Chemistry Courseware

David J. Yaron & Ryan P. Dwyer

Educational Data Mining

Online Learning

Integrated chemistry courseware such as REAL CHEM captures rich, fine-grained data on how students engage in the process of learning. Our early analyses show that students in the same course—and under identical policies—often navigate the materials in surprisingly different ways, with implications for both instruction and learning outcomes. To support instructors and researchers in interpreting these patterns, we developed an interactive visualization tool that displays each student’s sequence of interactions across units and modules, along with clustering methods that group students by behavioral similarity. Instructor use of the tool revealed substantial variation in engagement and highlighted cases where students were not using the courseware in the ways instructors intended. Some policies produced more consistent behavior, while others allowed wide divergence. These findings underscore the importance of using learning engineering to understand student pathways when designing courseware, interpreting learning data, and shaping course policies.

David J. Yaron¹ and Ryan P. Dwyer²

¹ Department of Chemistry, Carnegie Mellon University, Pittsburgh PA 15213; yaron@cmu.edu

² Department of Chemistry, University of Mount Union, Alliance, OH 44601; dwyerry@mountunion.edu

Abstract.

Keywords: Educational Data Mining, Online Learning

Introduction

REAL CHEM is an online courseware system that integrates instruction and practice for a two-semester general chemistry sequence. Unlike online homework systems, it replaces the textbook and practice tools with one environment that captures detailed traces of students’ process of learning [Koedinger et al. (2025)]. Early analyses show that students under the same policies often engage in very different ways. This project introduces a visualization tool for examining student pathways and clustering methods for identifying common engagement patterns. Framed as a nested learning engineering cycle, this work supports iterative, data-informed refinement of courseware design and instructional policies within a larger ongoing learning system [Craig et al. (2025)].

Approach / Design

REAL CHEM captures detailed logs of every student interaction. Our visualization maps these interactions by listing courseware opportunities vertically (organized by unit and module) and ordering each student's actions horizontally over time, with weekly boundaries for context. Symbols indicate interaction types, including formative (Learn By Doing, Did I Get This) and summative checkpoints. Hover actions reveal additional detail. Each pathway is also converted into a linear sequence suitable for clustering analyses.

Findings / Insights

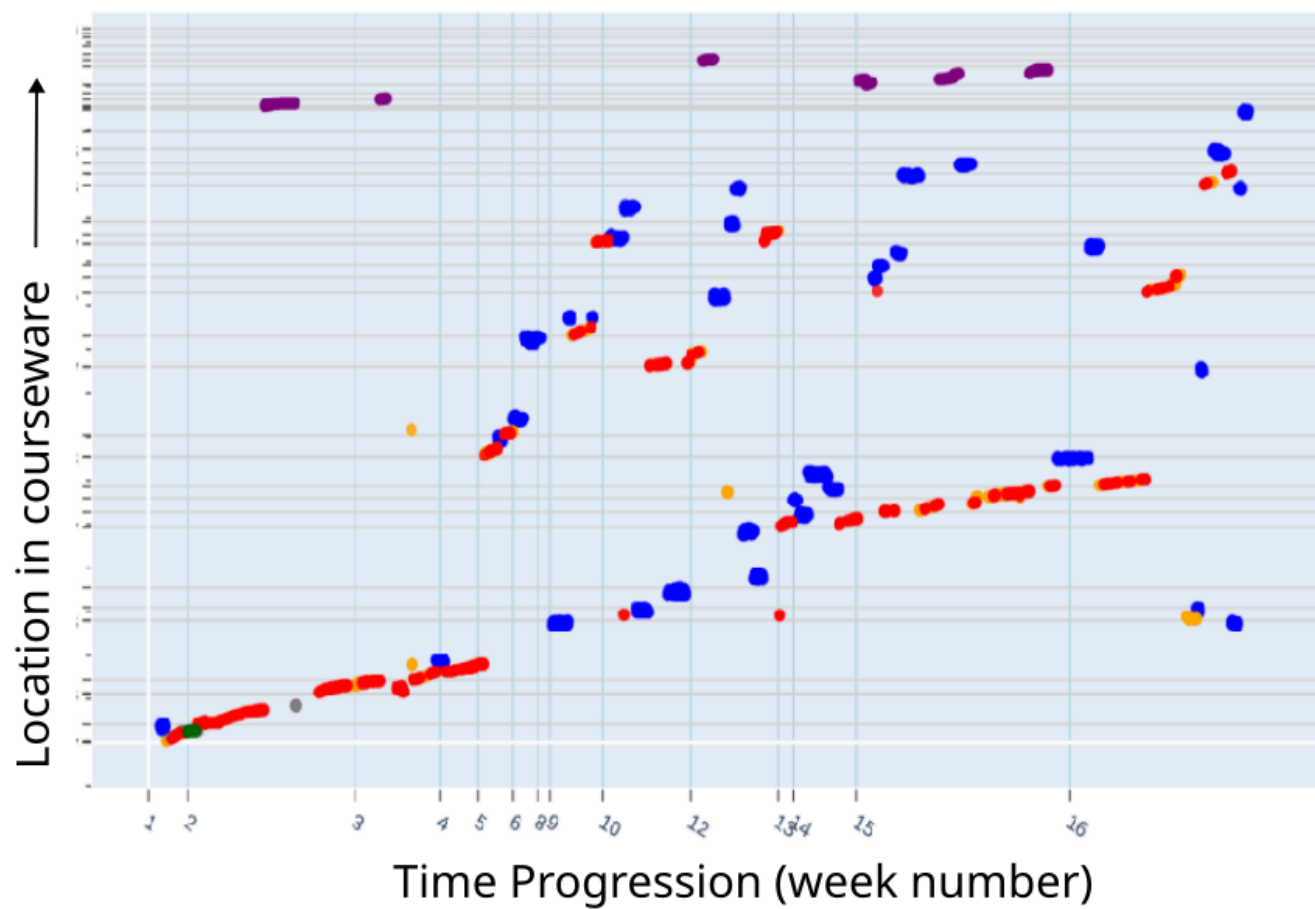
The tool revealed far greater variation in student behavior than expected, challenging design assumptions. Some course policies yielded more uniform engagement, but many instructors were surprised to see students using the courseware in ways that differed from their guidance. Clustering identified pathways ranging from linear progression to adaptive patterns in which students revisited formative practice based on prior performance. Further work is needed to refine these clusters and clarify their relationship to learning outcomes.

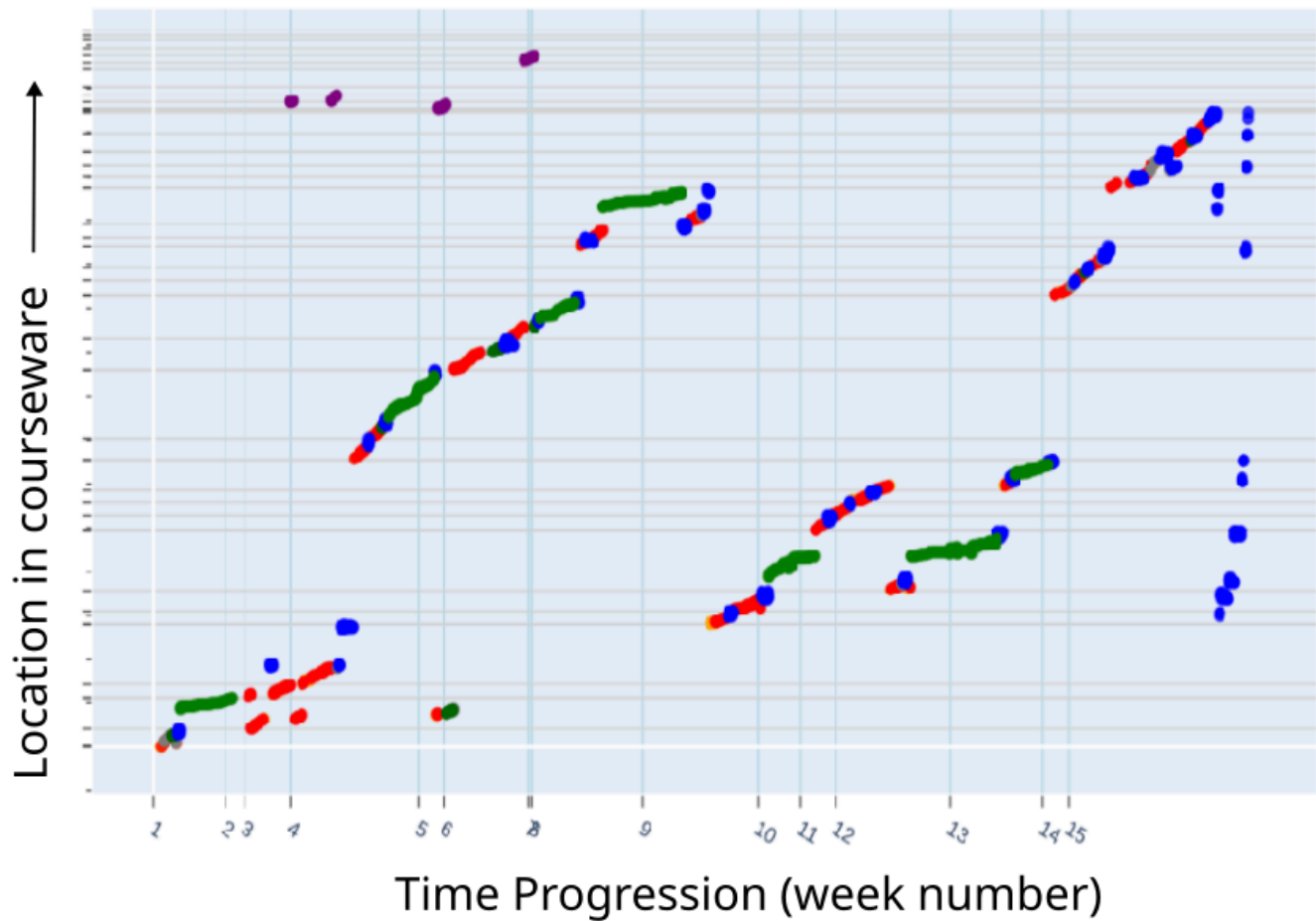
Implications

These findings illustrate the value of a learning engineering approach in which learner data is used to iteratively evaluate assumptions, refine design decisions, and adapt instructional supports. Recognizing this variation is important for designing supports, adaptive features, and analytics. Course policies can shape behavior, but their influence varies, underscoring the need for clearer guidance and tools that help instructors monitor and respond to engagement patterns.

Figure 1

Student pathway visualizations illustrate clear differences in engagement.





References

- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). <https://doi.org/10.59668/2109.21735>
- Koedinger, K.R., Blaser, M., McLaughlin, E.A., Cheng, H., Yaron, D.J. (2025) Journal of the American Chemical Society Au 5, (pp. 1268–1278). <https://doi.org/10.1021/jacsau.4c01126>



REAL CHEM Action Research Through the LearnLab Summer School

J. Bryan Henderson, Sarah M. Bass, Ryan P. Dwyer, Kimberly Larson, Chris Mead, Abigail Stein, & David J. Yaron

Action Research

Educational Technology

Professional Learning

REAL CHEM is a comprehensive courseware environment for general chemistry. Because the REAL CHEM courseware aligns with widely adopted textbooks and instructional practices, it has been implemented across a broad range of institutions. Given both the Learning Engineering and practitioner focus of this manuscript, we frame our method as a description of how REAL CHEM instructors were supported to conduct action research with curated data from their courses incorporating REAL CHEM. The corresponding results are the ongoing action research projects that have emerged. We highlight one of these projects and discuss implications for data-driven reforms of teaching practices as a valuable Learning Engineering cycle.

Introduction

REAL CHEM is a comprehensive courseware environment for general chemistry (Carnegie Mellon University, 2022) being iteratively refined through multiple cycles of Learning Engineering (Craig et al., 2025). It is built around the OpenStax Chemistry 2e textbook (Flowers et al., 2009) to ensure alignment with widely used open-source instructional materials, particularly in two-year institutions where consistent, scalable support is often lacking yet critical. Designed for hybrid use, REAL CHEM blends online and in-person instruction, integrating content and assessment into a coherent instructional thread that supports both teaching and learning. The platform places a strong emphasis on meeting the needs of at-risk students through scaffolded practice, timely feedback, and tools that help instructors target support. Every student interaction is instrumented to generate rich data that flows back to instructors to inform teaching, to developers to improve content, and to researchers working to advance the science of learning. In this way, REAL CHEM functions not just as a tool for teaching and learning, but as a continuously evolving platform for instructional improvement and research.

To engineer REAL CHEM as a learning experience that is accessible and usable by a growing community of education researchers, the technical infrastructure has been expanded with targeted support for instructors and researchers, including streamlined onboarding, implementation guides, and structured opportunities for collaboration. This includes the expansion of existing training programs with dedicated tracks to prepare faculty and early-career researchers to contribute effectively to courseware-based research. One such example is the support of action research by REAL CHEM instructors through the LearnLab Summer School.

Method

Given both the Learning Engineering (Craig et al., 2025) and practitioner focus of this manuscript, we frame our method as a description of how instructors were supported through the LearnLab Summer School to conduct action research with data from their courses that incorporate the REAL CHEM platform.

LearnLab Summer School

The LearnLab Summer School (LearnLab) is an intensive one-week program focused on supporting technology-enhanced learning environments and analyzing data generated in those environments (LearnLab, 2025). All LearnLab participants take part in daily whole-group activities where they attend talks on topics such as intelligent tutoring systems, computer-supported collaborative learning, and educational data mining. In addition to whole-group activities, LearnLab also offers programming in multiple subject tracks. REAL CHEM involvement in the 2025 LearnLab fell under the general chemistry subject track.

REAL CHEM Track-Specific Activities

Track-specific activities were supported by mentors involved in educational research surrounding REAL CHEM. Mentors would help instructors develop action research questions based on the available data and provide guidance for the corresponding data transformation and analysis necessary to pursue those questions.

Dataset Coordination and Emerging Action Research Questions

Each REAL CHEM instructor at LearnLab was provided a personalized dataset based on their students' engagement and performance within the REAL CHEM platform. Examples of action research based on this data included exploration of how different grading policies might associate with different levels of REAL CHEM student engagement, as well as how different paths by which students navigate through REAL CHEM potentially associate with chemistry exam performance.

Results

With supporting REAL CHEM instructors through LearnLab serving as the method of this Learning engineering and practitioner-focused manuscript, the corresponding results are the instructor projects exploring their action research questions that emerged from LearnLab. We highlight one of these projects now.

REAL CHEM Engagement and General Chemistry Scores

One REAL CHEM instructor utilized LearnLab to investigate how engagement with REAL CHEM is potentially associated with outcomes on the American Chemical Society (ACS) general chemistry exam (Bass, Carpenter, & Henderson, 2025). In large enrollment gateway courses (CHEM 101), the instructor examined whether progress completion, checkpoint assessments, semester timing, and prior knowledge influenced chemistry exam performance. Results showed that students who took the course in the spring (off-track) performed about 5.6 points lower than those in the fall (on-track). This finding suggests that targeted interventions could help close performance gaps between semesters and across student groups, and the instructor is now exploring possible interventions to reduce the gap their action research has identified.

Discussion

Supporting REAL CHEM instructors through LearnLab provides an example of how to make action research with big data more tenable for practitioners, both in terms of access to the data and mentor support with data transformation and analysis. While some of the LearnLab participants had limited experience with data-driven educational research, by the end of the week, multiple REAL CHEM instructors were analyzing learning growth curves, as well as conducting regression and principal component analyses on their personalized student data. Surveys of LearnLab participants suggest interest in conducting future data-driven reforms of their teaching. This continued interest is promising, and might not be the case if structured supports were not provided for data mining, transformation, and analysis. Curating big data for instructors and providing them personalized mentoring can potentially instill an improved sense of self-efficacy when it comes to their likelihood of engaging with data-driven instructional reform. This makes for a potentially valuable cycle of Learning Engineering (Craig et al., 2025).

Acknowledgments

We would like to acknowledge the organizers and participants of LearnLab for an engaging week of professional learning. We gratefully acknowledge the financial support from The Bill & Melinda Gates Foundation for this research effort. Opinions and conclusions expressed in this article do not necessarily reflect the views of this funding agency.

References

- Bass, S. M., Carpenter, T. S., & Henderson, J. B. (2025). Impact of Semester Timing on Student Outcomes in the First Course of a General Chemistry Sequence, presenter, 2025 Provost's Teaching and Learning Symposium, UMBC, September 2025.
- Carnegie Mellon University. (2022, June 8). Universities partner to make chemistry more equitable. <https://www.cmu.edu/news/stories/archives/2022/june/asu-cmu-chemistry.html>
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training

Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). <https://doi.org/10.59668/2109.21735>

Flowers, P., Theopold, K., Langley, R., & Robinson, W. (2019). Chemistry 2e. OpenStax.
<https://openstax.org/details/books/chemistry-2e>

LearnLab. (2025). LearnLab Summer School. <https://learnlab.org/learnlab-summer-school/>



SCALED: A Transferable Learning-Engineering Model for Large-enrollment STEM Courses

Medha Dalal, Jean Larson, Peter van Leusen, Scott Mahler, Ryan Meuth, & Phillip Miller

high-enrollment

multi-modality

Scalable course design

Extended Abstract

Introduction

High-enrollment gateway courses in engineering face pressures to expand access while improving quality and consistency across sections and modalities (Meuth et al., 2025). This poster presents Scalable Convergence Across Learning Experiences and Disciplines (SCALED) model that leverages learning engineering principles to address these pressures.

Our Approach: Learning-Engineering Intervention

We redesigned the first-year programming course as a unified “mega-course” that delivers a comparable learning experience across on-campus, online, and MOOC modalities. Using a flipped design, students progress through short videos, readings, low-stakes coding activities, tutorial-style labs, and weekly summative programming projects. Four proctored exams are mapped to micro-learning objectives to support comparability. Course materials are primarily delivered by zyBooks, a web-based interactive textbook platform, whose auto-grading and immediate feedback free the instructional team for coaching. Learning community support is integrated via Inscribe. Analytics flag students who spend more than 60 minutes on an assignment without improving, triggering outreach with hints and resources. Bi-weekly progress messages direct students to office hours and forums. These mechanisms sit within a nested Learning Engineering cycle (Craig et al., 2025) that analyzes assessment performance, time-on-task, DFW patterns, and end-of-course feedback for course revisions.

Findings: The SCALED Model

Over seven years, outcomes align with the model's aims. DFW rates have decreased, and retention increased relative to the lecture-centric format. End-of-course evaluations are reliably higher ($p < .014$ across seven large sections; $d > 1.75$). Enrolling 3,000+ students per semester across modalities, the mega-course has crystallized into SCALED, a technology-enhanced instructional model consonant with NCAT's principles (NCAT, n.d.) and characterized by a digital foundation (interactive content, auto-grading, analytics), collaborative design and delivery (coordinated teams of faculty, graduate and undergraduate TAs, graders), student-centered pedagogy (frequent feedback, active learning, learning communities), academic-unit coordination (program, advising, student support) and iterative outcomes-driven design.

References

- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). <https://doi.org/10.59668/2109.21735>
- Meuth, R., Dalal, M., & van Leusen, P. (2025). WIP: Scaling principles of programming. Paper presented at the IEEE Frontiers in Education (FIE) Conference, Nashville, TN.
- National Center for Academic Transformation. (n.d.). The National Center for Academic Transformation. <https://www.thencat.org/>



Terracotta: Lowering Barriers to Experimental Education Research

Benjamin Motz & Harmony Jankowski

Education

Experiment

Learning Management System

Expanded Abstract

Experimental research is the gold standard for establishing causal links between instructional practices and student outcomes. However, controlled experiments are rare in education due to their prohibitive cost, logistical complexity, and difficulty in naturalistic school settings. Meanwhile, the Learning Management System (LMS) has become central digital infrastructure in K-12 and higher education, directly shaping students' learning experiences (Davis et al., 2009; Hill, 2020; Lonn & Teasley, 2009; Mervosh & Swales, 2020; Pomerantz & Brooks, 2017). Current guidance for teachers on designing effective LMS assignments is minimal and often lacks a strong evidence base. We advocate for leveraging existing LMS infrastructure to lower the barriers to rigorous experimental research by manipulating LMS assignments. We present Terracotta, a novel, open-source LMS plugin that automates a complete, rigorous, and ethically responsible experimental research workflow (Motz et al., 2023).

Terracotta creates a seamless integration into existing LMS platforms using the Learning Tool Interoperability (LTI) standard, enabling experimental manipulation of assignment content and random assignment of students to different treatment conditions. The tool automates four requisite functions for conducting high-quality embedded experiments: random assignment, informed consent (with concealed student responses), mapping of outcomes to assignments, and de-identified data export. Terracotta supports various designs, including simple A/B tests and complex within-subject crossover designs (AB/BA) with pretest/posttest measures, which are otherwise notoriously difficult and expensive to implement manually.

Terracotta makes large-scale, individually randomized experiments feasible and cost effective within standard educational settings. By automating the most challenging logistical components, Terracotta bypasses difficulties that have hindered such studies in the past. The LMS provides a ready-made, large-scale platform for non-intrusive experimental manipulation of core instructional elements. While Terracotta has always enabled multiple choice, short answer, and file-upload assignments, now researchers can embed a Qualtrics survey (or any external web activity) in an LMS assignment and send personalized messages through the LMS (or via email). Automation drastically reduces the labor and errors associated with complex experimental designs like crossover studies, making them accessible to a wider research audience.

Using Terracotta, researchers can reverse the decline in experimental education studies while also casting wide nets to sample from more classes and institutions, enabling robust testing of the generalizability of findings across diverse student populations and pedagogical contexts (Slavin, 2002; Sullivan, 2011; Motz et al., 2025). Teachers and researchers can collaborate to embed research into routine learning activities, increasing teachers' involvement in research, and leading to more relevant insights about what works in practice. In total, the goal of Terracotta is to lower the barriers to easy, accessible, responsible, and rigorous experimental research across education levels, student populations, and learning materials.

References

- Davis, B., Carmean, C., & Wagner, E. D. (2009). *The evolution of the LMS: From management to learning*. Santa Rosa, CA: E-Learning Guild.
- Hill, P. (2020, April 3). Massive increase in LMS and synchronous video usage due to COVID-19. <https://philonedtech.com/massive-increase-in-lms-and-synchronous-video-usage-due-to-covid-19/>
- Lonn, S., & Teasley, S. D. (2009). Saving time or innovating practice: Investigating perceptions and uses of Learning Management Systems. *Computers & Education*, 53(3), 686–694.
- Mervosh, S., & Swales, V. (2020, March 12). Colleges and Universities Cancel Classes and Move Online Amid Coronavirus Fears. *The New York Times*.
- Motz, B., Üner, Ö, Jankowski, H., Christie, M., Burgas, K., del Blanco Orobítg, D., & McDaniel, M. (2023). Terracotta: A tool for conducting experimental research on student learning. *Behavior Research Methods*. doi: 10.3758/s13428-023-02164-8
- Motz, B., Chinni, A., de Leeuw, J., Jankowski, H., Aggarwal, A., Amato, M., Berlin, K., Britten, K., Brown, A., Cerchiaro, M., Evans, N., Findley, A., Gorman, R., Gregg, K., Hansen, K., Hollender, H., Hullinger, R., Larkin, P., Lion, M., Long, R., Mannarino, A., Mocko, M., Moore, K., Packowski, J., Palmer, C., Ritchie, K., Scott, J., Stanton, M., Talcott, L., Wagner, L., Waite, M., Yeager, R., and Fyfe, E. (2025). ManyClasses 2: The effects of prequestions on media interactions and learning, *Journal of Educational Psychology*. doi: 10.1037/edu0000978
- Pomerantz, J., & Brooks, D. C. (2017). *ECAR Study of Faculty and Information Technology, 2017*. EDUCAUSE.
- Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7), 15–21.
- Sullivan, G. M. (2011). Getting off the “gold standard”: Randomized controlled trials and education research. *Journal of Graduate Medical Education*, 3(3), 285–289.







The Promise of Scenario-Based Assessments for College Instruction

Jonathan D. Cohen, Omer Ari, Emily Bryans Dobar, Gertrude Tinker Sachs, John Sabatini, Tenaha O'Reilly, & Joseph P. Magliano

Assessment-based Learning

Higher Education

scenario-based assessment

The lecture-high stakes test structure of many college courses does not promote deep learning that can be transferred to real world contexts. Scenario-based assessments (SBAs) provide an approach to assessment that grounds authentic problem solving and scaffolds students to a solution for those problems. The paper reports the first iteration of a design-based research (DBR) design cycle to develop an SBA authoring system, which will enable college instructors to use SBAs in their courses. An SBA was administered in an introduction to interdisciplinarity course taught at a large, public university. Data from students and faculty suggest that it was usable, and they perceived it to support learning beyond traditional course assessment formats. We reflect upon the affordances of SBAs that led to these positive perceptions.

Introduction

Many college courses follow a structure that is arguably not optimal with respect to promoting deep learning that can transfer beyond the classroom. Instructors in many disciplinary-specific courses (e.g., psychology, biology, history, etc.) structure their courses around a series of lectures interspersed with high-stakes assessments. Moreover, many of these tests adopt formats (e.g., multiple choice tests), that are problematic because they (a) are not grounded in theories of learning (Mislevy & Yan, 2017), (b) are decontextualized (O'Reilly et al., 2018), and (c) discourage transfer (Foster & Piacentini, 2023).

Students deserve better. Scenario-based assessments (SBAs) provide a potential solution to this systemic dilemma (Sabatini et al., 2020). SBAs are designed to afford students opportunities to apply the knowledge and skills they acquire in their courses. They ask students to address or solve broad, purpose-driven problems in ecologically valid settings. A well-designed SBA walks students through a task sequence that models strategies used by experts, while providing valid and reliable measurement of critical skills. SBAs typically have the following characteristics (e.g., Sabatini et al., 2020): (a) they explicitly model the critical steps in a complex performance task; (b) they ground that task in the context of solving a real-world problem; (c) the task sequence simulates task-appropriate strategies, improving student engagement and providing opportunities to develop critical metacognition and self-regulation skills; and (d) individual tasks (or items) measure critical supporting skills (or key supporting knowledge) needed to carry out specific steps in the task sequence.

The current study is the first iteration of a design-based research (DBR) design cycle (Reeves, 2006) intended to lead to an SBA authoring system for college instructors. DBR parallels learning engineering (Baker et al., 2022) in that each involves iterative, rigorous, data-driven analysis of theoretically informed interventions to support the development of learning environments. We designed and implemented an exemplar SBA in order to better understand both the process of designing SBAs and the role the SBAs played in the students' learning. It was guided by the following research questions: (a) What was the perceived usability of the SBA? (b) What are the student perceptions of the SBA? and (c) What are the instructor's perceptions of the SBA? To answer these research questions, we employed a concurrent mixed methods design, leveraging data collected from students and the instructor of an introduction to interdisciplinary thinking course in Spring 2025.

Methods

Participants, Course Context, and the Scenario-Based Assessment

Nine students enrolled in an introduction to interdisciplinarity course at a large, public university participated in the study. The university is an urban, primarily minority serving institution. The participants resembled the university's students in general, with most students (66%) between 18 and 24 years old. Forty-eight percent of university students identify as Black or African American, 22% as White, 21% as Hispanic. Nearly half (46%) of students are first-generation college students.

Interdisciplinarity in Human Learning and Development is a course in which students learn about the concept of interdisciplinarity and interdisciplinary thinking. Repko et al.'s (2019) textbook, *Introduction to Interdisciplinary Studies*, is used as the core reading. It introduces students to the foundational knowledge of interdisciplinarity, then it introduces the Broad Model of the Interdisciplinary Research Process (IRP). The Broad Model involves six steps, from defining the problem to a literature search, and reflection on the problem. The course culminates with a paper in which students apply the model to a societal problem of significant personal interest, following the criteria specified for each of the steps.

The SBA situates students in a scenario in which they are college-aged interns at Comm360, a fictional community-based nonprofit. They are part of a team of interns, tasked with developing a structure for a report to the CEO of Comm360 describing how the nonprofit might address the issue of high school dropouts. The digitally delivered SBA is broken into screens that parallel the Broad Model. As students progress through the screens, they participate in activities such as

highlighting tasks, article summarization, and other constructed responses. The students' work follows them through the SBA, so their responses on activities form the foundation for subsequent activities.

Procedure

We collected data from the following sources: (a) a usability questionnaire delivered following the SBA implementation; (b) student artifacts, including their interactions with and responses to the SBA and reflections on the SBA process; and (c) instructor interviews.

We employed reflexive thematic analysis to analyze the qualitative data (Braun & Clarke, 2006), in which we collaborated through the process, developing a shared interpretation of the data. For this study, a group of five researchers, all of whom are grant personnel and including the course's instructor, participated in the analysis. The first step was familiarization with the data, followed by the first phase of inductive coding. From these codes, we generated a list of themes, which we subsequently discussed, refined, and named. These themes focused our continued exploration of the data. We used basic descriptive statistics to analyze the usability questionnaire data. The questionnaire is a modified version of the System Usability Scale (Brooke, 1996), which is a 10-item questionnaire, widely used for this type of evaluation. Our modified version adapted 9 of the items for our study. It has the advantages of being brief and can be used with small sample sizes, but lacks in fine-grained, formative diagnosis of specific design issues.

Results

What Is the Perceived Usability of the SBA?

Of the consenting students, 11 completed the usability questionnaire. Participants rated 8 of the 9 survey items at or above a 4.00 on a 5-point Likert scale (1, Strongly Disagree to 5, Strongly Agree), suggesting that, overall, participants felt positively about the usability of the SBA. Most of the items explored students' perceived usability of and confidence with the SBA interface, but the first item was I felt confident in my understanding of the content to gauge students' comfort with the content of the SBA. This item received an average rating of 4.18, suggesting 10 of the 11 students felt confident with the content. With an average rating of 4.27, 10 of the 11 participants also indicated agreement with I found the content to be consistent with what we were learning in the course.

The highest-rated item was I think that I would need the support of a technical person to be able to use the SBA interface. After reverse-coding this item since it was negatively worded, the average rating was 4.64 out of 5 (1.91 before reverse-coding), suggesting 9 of the 11 students indicated they could easily use the interface without external support. The lowest-rated item received an average rating of 3.73, I think that I would like more SBAs to be used in other courses, a result which will be explored in future iterations of the SBA.

What Are the Student Perceptions of the SBA?

An overarching theme derived from the students' reflections is that their experience applying the Broad Model was much messier than they anticipated. The weeks leading up to their participation in the SBA were spent learning the model as linear, sequential steps in a process, as is introduced in the textbook. The SBA, though, simulated the application of the Broad Model in an authentic setting, and the nature of that simulation required a more recursive application of the Broad Model. This shift in understanding was a revelatory moment for some of the students. Some correctly perceived that their previous conceptions of the model as linear were more a function of its presentation in the textbook, and their reflections demonstrated an understanding of the need to modify the process to react to circumstance. For example, one step of the model involves explicitly justifying an interdisciplinary approach. The scenario they were presented with elided this step, as one student

described: “In the SBA we never really had to do that, our ‘supervisor’ was immediately on board.” Other students experienced the recursive application of the model similarly, noting that this was a change in their previous understanding of the model.

Another experience noted by the students was their interactions with their simulated colleagues (i.e., dialogic agents) within the SBA. The simulation was minimal; the “colleagues” were simply images and text, and there was no real interactivity between them and the students. However, some students reflected that even that minimal level of simulation of collaboration supported their work. This was evident in the use of the first-person plural in many of the students’ reflections when referencing their action steps during the SBA (e.g., “our discussion”, “our team at the non-profit”, etc.). Some students explicitly referenced the agents as important parts of the learning experience. One noted that the introductions to their “colleagues” at the beginning of the SBA “made me think it was real classmates”. This connection appeared to be a motivating factor for some of the students, who referenced the “realness” of the teamwork when describing completing the SBA.

In addition to the motivational aspect of the agents, some students reflected on how the scaffolding from these digital teammates were important parts of their SBA experience. For instance, one remarked that “having the other interns’ answers available allowed me to check and compare my answers.” Another discussed how the agents’ perspectives “broaden[ed] my horizons past where I would’ve originally looked at.” Clearly the students did not believe their SBA colleagues were actual people, but the students seemed to suspend disbelief while working through the scenario, allowing for a more “real” seeming environment. This realness was a constant theme in the students’ reflections.

What Are the Instructor Perceptions of the SBA?

The instructor for this course has seven years of experience teaching the Broad Model. Watching his students use the model over the years, the instructor noted that the step-by-step application of the model often fell short of the potential for promoting interdisciplinary thinking. He observed that students tended to check off steps performatively. Some of this is a consequence of the prescriptive nature of some steps of the model. For example, one step asks students to declare that the problem is complex, resulting in students simply writing “The problem is complex.” The order of some steps also did not make sense. In one of the earliest steps, students were asked to state that there was more than one discipline offering insights into the problem before they had actually reviewed the literature and confirmed that was the case. Implementing the IRP steps in this format felt like busywork, leaving students without the rich insights they were supposed to gain from following the model. He characterized the work as often not “cognitively engaged” and that it felt “uninspired.”

This contrasted with students’ performance on the SBA, which the instructor described as richer and deeper. As he put it, “the SBA brought the Broad Model to life,” as the steps were applied authentically within the simulated context of a nonprofit committee tasked with developing a more comprehensive understanding of the causes of the high school dropout problem in the community they serve. Embedded in the natural-sounding dialogic interactions, the steps were applied purposefully and authentically to the work of the committee. Students were more engaged in this more fluid interdisciplinary thinking process, which incorporated the same six steps of the Broad Model of IRP. He indicated that students were more invested in the topic and wanted to continue working towards a solution, with one student commenting, “We evaluated different perspectives but never got to the problem-solving portion.”

Discussion

This study demonstrated that students found the SBA usable, and in general, the students’ reflections and the instructor’s perceptions were consistent in reporting that students developed a more nuanced and realistic understanding of how the Broad Model can be used to solve problems that are interdisciplinary in nature. The SBA format afforded the designers opportunities to create a structure which helped students achieve the learning objectives. In particular, a primary affordance of the SBA is its ability to approximate an authentic setting within which to situate students’ work. Certainly, part of that authenticity stems from having images to represent the agents and including language that sounds truer to the types of

individuals students could reasonably expect to encounter in the scenario. Beyond these more superficial elements, the authenticity stems from situating the SBA in a real-world environment that demands fluidity and complexity. The way models are presented in textbooks are flat representations, which push students toward a flatter understanding of processes. SBAs afford situated application of knowledge and skills rather than simple recitation and therefore help students to develop a more nuanced understanding of those knowledge and skills (e.g., Bransford, 1997).

The SBA format also affords designers the ability to model behaviors and skills related to the learning objectives. In this SBA, the interactions between the students and the dialogic agents provided opportunities to model the application of the steps of the Broad Model. Research demonstrates that dialogic agents support learning, even in a context when they are not truly interactive with students (Discoll et al., 2003). It was through the modeling and scaffolding provided through the dialogic agents that students were led to understand the steps of the Broad Model are less linear and more recursive than they perceived them to be after reading the textbook.

It is also notable that this SBA was implemented in an asynchronous online environment. Such environments can feel isolating for students (Croft et al., 2010), as student-student and student-instructor interactions are digitally mediated and temporally separated in asynchronous modalities. The presence of agents—even minimally interactive ones, as found in this SBA—was a motivating factor for the students.

Scenarios form the backbone of SBAs and afford designers the ability to integrate sequenced activities which can scaffold cognitive processes. At the beginning of this SBA, the scenario provided an authentic opportunity to activate students' prior knowledge. Later parts of the SBA attached targeted literacy activities (i.e., finding and highlighting relevant text) to the narrative. The SBA concluded with activities asking the students to synthesize ideas they generated throughout the SBA. In real-world situations where students actually work at a community non-profit, they could reasonably expect to do similar activities, adding to the perceived authenticity of the experience. Moreover, the structure of the scenario narrative forces designers to consider a range of pedagogical moves and scaffolding that fit into the narrative, as the scenario will feel flat and inauthentic if they do not.

In conclusion, this study suggests that SBAs can be applied to post-secondary instructional contexts. Their affordances may allow college instructors to break from the traditional lecture-high states exam structure of college courses and develop a context that can help students learn that they can transfer what they learn to real-world problems.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240021 to The University of Memphis. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*, 3(1), 1–23. <https://doi.org/10.1037/tmb0000058>
- Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, 8(2–3), 70–91. <https://doi.org/10.1080/15366367.2010.508686>
- Bransford, J. D. (1997). *The Jasper project: Lessons in curriculum, instruction, assessment, and professional development*. Routledge.

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.
- Brooke, J. (1996). SUS: a “quick and dirty” usability scale. In P. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189–195). CRC Press.
- Croft, N. Dalton, A. & Grant, M. (2010) Overcoming isolation in distance learning: Building a learning community through time and space. *Journal for Education in the Built Environment*, 5:1, 27-64, DOI: 10.11120/jebe.2010.05010027
- Driscoll, D. M., Craig, S. D., Gholson, B., Ventura, M., Hu, X., & Graesser, A. C. (2003-2004). Vicarious learning: Effects of overhearing dialog and monologue-like discourse in a virtual tutoring session. *Journal of Educational Computing Research*, 29(4), 431–450. <https://doi.org/10.2190/Q8CM-FH7L-6HJU-DT9W>
- Foster, N., & Piacentini, M. (2023). *Innovating Assessments to Measure and Support Complex Skills*. Organisation for Economic Co-operation and Development.
- Mislevy, R. J., & Yan, D. (2017). Evidence-centered assessment design and probability-based inference to support the generalized intelligent framework for tutoring (GIFT). In R. A. Sottolare, A. C. Graesser, X. Hu, & G. A. Goodwin (Eds.), *Design Recommendations for Intelligent Tutoring Systems: Volume 5—Assessment Methods* (Vol. 5, pp. 101–124). U.S. Army Research Laboratory.
- O'Reilly, T., Sabatini, J., & Wang, Z. (2018). Using scenario-based assessments to measure deep learning. In K. K. Millis, D. Long, J. P. Magliano, & K. Wiemer (Eds.), *Deep Comprehension: Multi-disciplinary Approaches to Understanding, Enhancing, and Measuring Comprehension*. (1st ed., pp. 197–208). Routledge.
- Reeves, T. (2006). Design research from a technology perspective. In J. Van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational Design Research*. Routledge.
- Repko, A. F., Szostak, R., & Buchberger, M. P. (2019). *Introduction to Interdisciplinary Studies*. SAGE Publications.
- Sabatini, J., O'Reilly, T., Weeks, J., & Wang, Z. (2020). Engineering a twenty-first century reading comprehension assessment system utilizing scenario-based assessment techniques. *International Journal of Testing*, 20(1), 1–23. <https://doi.org/10.1080/15305058.2018.1551224>



“Walk It Out”: An Embodied and Mobile AI Tutor for STEM Education

Mina C. Johnson-Glenberg, Feni Vishnubhai Chaudhari, Srujal Jitendrakumar Patel, Sophia Baia, Rebecca Vieyra, Daniel O'Brien, & Colleen Megowan-Romanowicz

Embodied Learning

Embodied STEM

Instructional AI

LLM AI Tutors

Mobile Learning

Socratic Tutors

Interpreting motion graphs can be challenging for students. Classroom teachers lack the capacity to provide rapid, personalized feedback for each student during mobile learning activities. We created an embodied smartphone app using LiDAR technology that has been shown to significantly improve kinematics understanding. However, the app's current text-based feedback can be enhanced with an AI tutor that gives more adaptive and "Socratic" (querying and non-directive) feedback. This article describes the design process for creating a Socratic-style mobile AI tutor and our Retrieval Augmented Generation (RAG) LLM architecture. The RAG pulls from the authors' published base of physics education articles to ground the model. Human experts determined that Claude Sonnet 4.5 provided significantly more precise, reliable, and Socratic-style feedback compared to other LLMs. This team of learning engineers and physics teachers is now creating a semantic benchmark test to assess the quality of the LLM's feedback.

Many students struggle with making and interpreting motion graphs (Börner et al., 2015; 2019); Börner et al. describe the broader challenges in data-visualization literacy across many age groups. Our research group has created a mobile app that helps students interpret graphs in an embodied manner by "walking the graph" (Megowan-Romanowicz et al., 2022). Mobile learning (Sharples & Pea, 2014) has grown rapidly with the widespread classroom use of smartphones and tablets. Although sonic motion sensors have been used by physics teachers since the late 1980s (Brasell, 1987), the corrective feedback is usually not adaptive nor personalized. The lessons are often not embodied and the feedback is usually restricted to students self-interpreting how well their motion visually matched a pre-determined motion graph.

Newer phone and tablet devices enable learning to occur "anywhere, anytime", and use built-in sensors to capture real-world physical actions, such as motion or precise location (Huang & Chiu, 2015; O'Brien, 2021). Emerging AI-driven educational apps aim to address the issue of giving more personalized feedback, but their performance is inconsistent, particularly in real time (Guo et al., 2025; Kestin, 2025). There remains a clear need for Intelligent Tutoring Systems (ITS) that combine mobile sensor data with adaptive, embodied, and pedagogically sound AI feedback to enhance students' ability to make sense of graphical data representations. To our knowledge, no prior research has combined mobile LiDAR-based motion tracking with real-time AI feedback generated by large language models (LLMs) on graphing tasks. The current app is called Motion Visualizer (MV); it uses students' real-time locomotion (walking) to create immediate position-time graphs on a LiDAR-enabled smartphone or tablet. Our research questions center around creating an AI tutor and assessing the efficacy of the tutor designed to give Socratic-style, non-directive feedback. The work adheres to best practices in learning engineering (Goodell & Kolodner, 2023). That is, we use team-based designs, iterative processes, and team-based refinement consensus to ensure the AI feedback is concise, accurate, and helpful for students.

Research Questions:

RQ1) Using numerical inputs (embodied position and time locomotion data) and students' think-aloud language, is it possible for a symbolic language-based LLM to generate feedback responses that are accurate and helpful?

RQ2) How can the LLM prompts be optimized to support some of the best practices of pedagogy, i.e., Socratic questioning and adaptivity to students' ongoing errors?

RQ3) Is the field in need of new semantic benchmarks for evaluating AI tutor's graphing feedback?

Methods

The App

The Motion Visualizer (MV) app is one within a suite of apps in the Physics Toolbox Sensor Suite mobile application (Vieyra, n.d.). The MV tool was created with the support of a National Science Foundation grant and was launched to the public in 2023. The app was designed to integrate principles of embodiment and locomotion into STEM education. It uses Light Detection and Ranging (LiDAR) to calculate the distance from a stationary surface, such as a wall, to the phone. On the device screen, the student's position is plotted every 100 milliseconds, i.e., how far the learner is from the wall or stationary surface. As learners move toward and away from the wall, they can see their motion plotted on a graph almost instantaneously. See Figure 1 for a screenshot of a successful attempt at matching the purple target curve. When learners match their motion plot to the purple curve with an MSE of less than .10, they are rewarded with digital confetti.

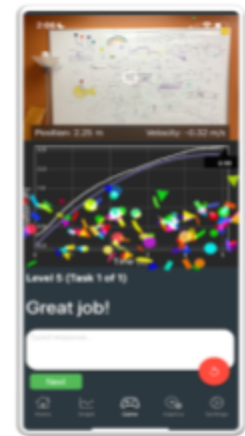


Figure 1. Screenshot.

Figure 2 shows an example of a failed attempt to match the purple target curve. The white dotted curve represents the student's locomotion. The non-AI version of MV gives only two types of text-based feedback: "Great Job" or "Try again". However, in our observations of teachers and peers supporting each other as learners, we noted the kind of feedback the students most benefited from was more nuanced. Effective feedback included suggestions about focusing attention on the starting position, considering how the target graph curves, and how the target graph shape might suggest how the learners' motions must change over time. Ultimately, the best kind of support helped students to not only build intuition about the meaning of position-time graphs, but to explicitly reflect on what the graph was telling them about their motion.

Data to test the AI models was gathered during a recent study

(Johnson-Glenberg et al., in preparation); the test set includes 90 transcripts of individual students' think-alouds and logfiles of the devices' "x, y" position-time coordinates across 14 graphing tasks. The tasks increased in difficulty over time; the final task was to recreate and match a vertical up parabola over a duration of 10 seconds.

What does it mean to be Socratic?

A difficulty in creating a pedagogically sound AI tutor is that LLMs have been pre-trained, hence predisposed, to give students the correct answer immediately. One of our goals was to create a feedback-giving system that would prompt the human user with a series of thought-provoking questions to encourage deep learning. The Socratic method of instruction (Paul & Elder,

2007) focuses on guiding learners to discover answers themselves via deduction and through the process of completing 14 graph-matching tasks. Directive feedback was given when necessary.

Comparative LLMs

We probed three popular LLMs: Claude Sonnet 3.7, Claude Sonnet 4.5, and Meta Llama 3.1 70B, using the same prompts and RAG. Three human experts analyzed the results and came to a consensus that the most precise, consistent, and concise results were achieved by Claude Sonnet Model 4.5. Thus, Sonnet 4.5 was chosen as the primary model. This exercise alerted us that the field needs a new type of automated benchmark, one that is sensitive to physics feedback that will take into account semantics and the embodiment that goes into walking and creating digitized graphs (an admittedly new field of study). Updates on the new benchmark creation using an SLM and vectorized idea units are forthcoming.

Our AI Model

The goal now was to optimize the Sonnet Model 4.5 into an embodied Socratic Feedback Model. There are two levels of prompts. The first is called the global prompt, which sets the overarching principles that the AI system should adhere to for feedback across all responses. Our global prompt clearly identifies the role for the LLM: "...act as a professor of physics who adopts the Socratic method of instruction by using guiding questions that do not provide direct results of the solution to the problem faced by the learner during the first seven attempts". The consensus among two of the physics teachers in our research team was that they, as human tutors, would only give a student more directive feedback after the student had failed seven times. Based on their experience observing students, they would give explicit feedback only after failure seven times in a row, as students often began to show signs of frustration after that. The global prompt also defines the various forms of common mistakes that learners might make during the tasks. These common errors included having the wrong starting point of the graph and producing graphs whose slopes were too shallow or too steep (greater than .10 MSE from target line). After analyzing too many chatty and wordy LLM feedback responses (e.g., the early Claude responses always started with, "I noticed that..."), the training team decided to constrain the AI system to not give feedback that would surpass a 35-word limit. This also takes into account the UI space constraints on a smartphone.

In a parallel fashion, the system incorporated level-specific prompts that further refined the global prompt. "Level-specific" means the feedback was yoked to each unique instance of the 14 graphs. These level prompts enabled the LLM to make decisions on whether to 1) prioritize start position state, 2) the dynamics of the slope, 3) changes in direction, or 4) learners' acceleration patterns. The more advanced the task levels, the more complex the model's reasoning became, given that it must simultaneously incorporate Socratic questioning. Figure 3 shows the architecture of the AI model. The LLM tracks the number of errors, and, after the 7th attempt, the feedback becomes more directive. This was achieved by including the following language in the global prompt: (Note: Capitalization increases the likelihood that a Claude model will give special emphasis to the instructions.)

"You guide students in a SOCRATIC manner. This means you ask questions until they have made 7 attempts. On the 8th attempt, you can start to give the students explicit answers that do not have a QUESTION MARK, but tell them exactly what they are doing wrong. The students' goal is to match the TARGET LINE."

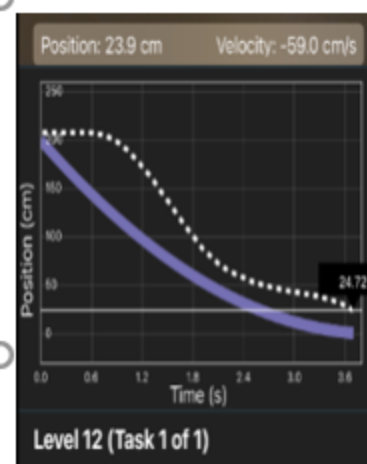


Figure 2. *Level 12, the purple line is the line the learners need to walk and match.*

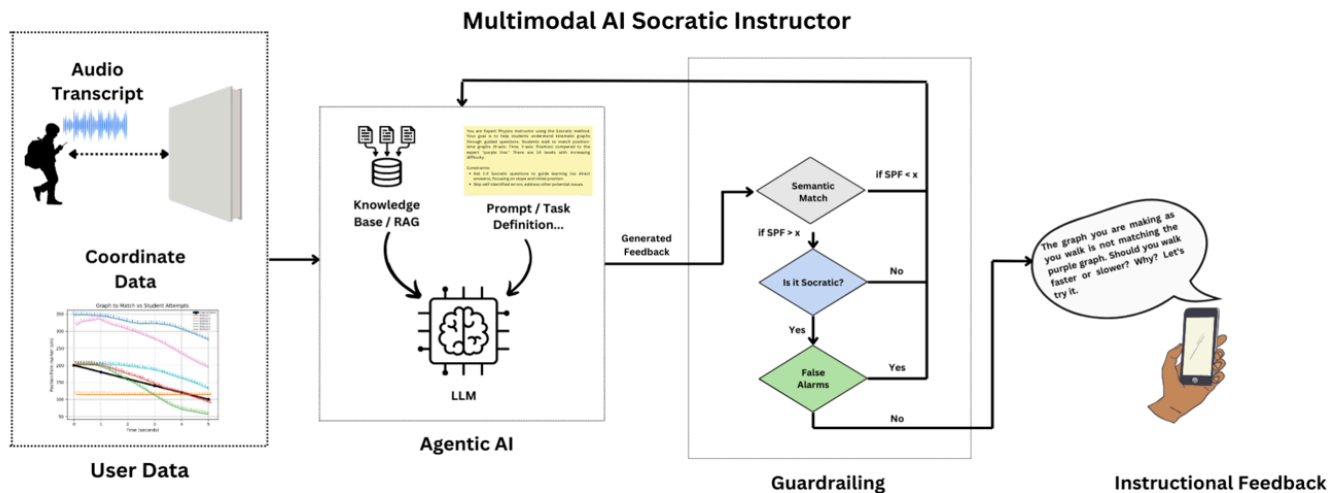


Figure. 3. Architecture of our Socratic AI Model.

Retrieval-Augmented Generation

To ensure that feedback is informed by studies on physics education, we created a lightweight retrieval augmented generation (RAG) module comprised of three papers on motion graph learning, education using smartphone sensor data, and previous studies from the design team (O'Brien et al., 2023; Vieyra et al., 2023; Vieyra et al., 2024). The RAG broke the papers into sections, and created a small vector index into which the sections, or chunks, were placed. For every attempt by the students, the RAG identifies types of errors and searches for extracted passages in the articles to inform its output. The passages are not directly exposed to or visible to the students, but serve as context for us, the testers, to ensure outputs are informed by solid kinematics research.

Results

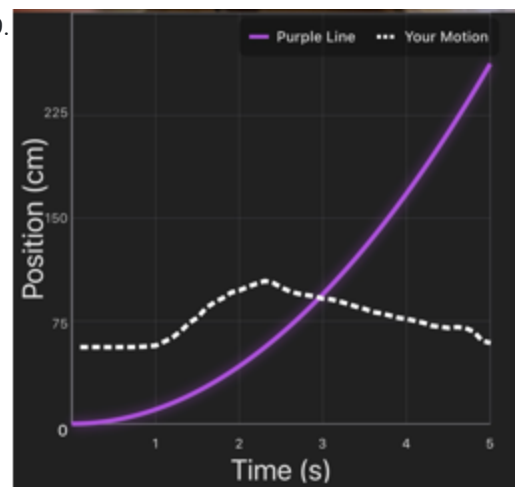
To analyze the quality of feedback produced by the AI system, we assessed a corpus of 660 student attempts to match the graph across all 14 of the available tasks in MV. For each attempt, we determined if the output of the AI system satisfied the following criteria, including if it: 1) pointed to the correct errors, 2) adhered to the constraints of the Socratic querying syntax, and 3) suggested fixes that corresponded to the student-initiated locomotion. Below are examples of correct and incorrect feedback:

Correct Feedback Example: Figure 4 illustrates the student's attempt at Level 9. The initial discrepancy in this attempt is with the learner's start position.

Model Feedback Attempt #1: "Where should you be at the start so your line begins near the target line?"

After multiple attempts, the student is now on their fourth attempt (Figure 4), and the model provides the following feedback.

Model Feedback #4: "Note your starting position. Did you start too far or close? Should you speed up in the beginning or slow down to better match the purple target line?"



This feedback was considered effective as it maintains the required order of error-priority within the system, first nudging the learner to think about their mistake in the initial position, then moving towards analyzing higher-order thinking concerning adjusting the slope.

Incorrect Feedback Example: In an earlier training epoch, the AI feedback was:

Model Feedback Attempt #2. “Your line stays above the target line for most of the walk. What if you started moving faster earlier? Would that help your line drop to match the target’s steeper curve?”

This feedback is incorrect as it indicates that the student’s line is consistently above the target curve, and moving faster would not decrease the slope or make it “drop”.

F1 Score Evaluation

To provide an initial quantitative check on model performance, an F1 score was calculated using 50 randomly selected feedback samples. Using our classifications of each response as appropriate or not, the model achieved an F1 score of 0.81, suggesting strong alignment with the intended error-detection and feedback goals. See a confusion matrix in Figure 5. We are currently analyzing the responses within the dark blue cell (correct rejections) with a sensitivity analysis on the type of error and appropriateness of response with a new pilot semantic benchmark.



Figure 5. Confusion Matrix

Discussion

We were consistently surprised by the time it took to train the LLM models to be precise, non-directive, and concise. We found that transcribing students’ vocal performance of their think-aloud protocols and then adding those transcripts as inputs to the LLM model was not helpful. This is because students say phrases like, “I need to go up”. The non-embodied LLM does not know the student is thinking of himself/herself as the line itself. It does not understand transmogrification or bodily metaphors.

Even after withholding the verbal inputs and feeding it only the position-time data, the models still ‘hallucinated’ on occasion and made incorrect analogies e.g., “imagine you are a ball rolling down a hill”. This type of AI-tutor feedback is especially egregious because one persistent error in understanding graphs is that naive human learners treat graphs like pictures (Beichner, 1994). After 12 months, and numerous LLM upgrades, the acceptability proportion of the output is now close to .81. The end goal is to encourage back-and-forth, multi-turn discourse between the tutor and the student. We are closer to that now that the model has been trained to remember and integrate the student’s previous trial performance ($n = 3$). We are examining recursion for memory using a model based on hidden units and working memory (Johnson-Glenberg, 2000). Our new system creates a lightweight memory for each student by storing key features of every attempt, including specific errors detected, slope issues, and starting-point mismatches. When a student submits a new trial, the model retrieves the last three attempts, summarizes common patterns of mistakes, and places this summary at the top of the level prompt used to generate the new feedback.

Future and Conclusions. We have created a Socratic AI tutor for a highly embodied graphing task with a LiDAR-enhanced mobile device. The model is adequate in theory, based on our evaluation of the model’s output, but still needs refinement and optimization. We will start to test its efficacy in the next semester with college students. Some researchers argue that because the current LLMs are trained solely on symbolic text that is inherently disembodied (i.e., culled from webpages and not experiential) and the models do not include perception, that the models will continue to make incorrect assumptions and never

be able to ever fully “reason”. Rodriguez (2025) argues, “...that the path to genuine AGI necessitates a foundational shift towards embodied experience, proposing that intelligence is inextricably linked to the perceptual and action-oriented constraints of a physical body interacting with a persistent world.” We concur. The plan is to push on the boundary conditions using symbolic text to help understand how to give relevant feedback for our highly embodied locomotive task. The next iteration may include perceptual inputs, something others have argued for, although we remind readers that using the think-alouds led the models to sometimes go astray. Thus, being more multi-modal did not help in this constrained physics task. The Socratic AI Tutor is currently running with AWS. A newly proposed upcoming research study will ascertain learning gains between the student group that uses the non-AI format of feedback in place now compared to the group that uses the Socratic AI tutor.

Acknowledgement

This material is based upon work supported by the National Science Foundation grants 2114586 and NSF NAIRR Supplement 2439960.

References

- Beichner, R.J. (1994). Testing Student Interpretation of graphs. *Am. J. Phys.* 62, 750–762 <https://doi.org/10.1119/1.17449>
- Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences*, 116(6), 1857–1864. <https://doi.org/10.1073/pnas.1807180116>
- Brasell, H. 1987. “The Effect of Real-Time Laboratory Graphing on Learning Graphic Representations of Distance and Velocity.” *Journal of Research in Science Teaching*, 24 (4): 385–395. <https://doi.org/10.1002/tea.3660240409>.
- Goodell, J. & Kolodner, J. (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Routledge. <https://library.oapen.org/handle/20.500.12657/86250>
- Guo, S., Halim, H. B. A., & Saad, M. R. B. M. (2025). Leveraging AI-enabled mobile learning platforms to enhance the effectiveness of English teaching in universities. *Scientific Reports*, 15(1), 15873. <https://doi.org/10.1038/s41598-025-00801-0>
- Huang, Y.-M., & Chiu, P.-S. (2015). The effectiveness of a meaningful learning-based evaluation model for context-aware mobile learning. *British Journal of Educational Technology*, 46(2), 437–447. <https://doi.org/10.1111/bjet.12147>
- Johnson-Glenberg, M. C. (2008). Fragile X syndrome: Neural network models of sequencing and memory. *Cognitive Systems Research*. 9 (4): 274–292. doi:10.1016/j.cogsys.2008.02.002.
- Kestin, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15(1), 17458. <https://doi.org/10.1038/s41598-025-97652-6>
- O’Brien, D. (2021). A guide for incorporating e-teaching of physics in a post-COVID world. *American Journal of Physics*, 89(4), 403–412. <https://doi.org/10.1119/10.0002437>
- O’Brien, D., Vieyra, R., Cortés, C., Johnson-Glenberg, M., & Megowan-Romanowicz, C. (2023). Evaluating learning of motion graphs with a LiDAR-based smartphone application. *arXiv:2301.10334*. <https://doi.org/10.48550/arXiv.2301.10334>

Paul, R. & Elder, L. (2007). *The Art of Socratic Questioning*. Dillon Beach, CA: The Foundation for Critical Thinking.

Rodríguez, E. Q. (2025). The body is the key: A formal architecture for embodied grounding in artificial general intelligence (Introducing the virtual embodiment module. <https://doi.org/10.13140/RG.2.2.11025.70244>

Sharples, M., & Pea, R. (2014). Mobile learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 501–521). Cambridge University Press. <https://doi.org/10.1017/CBO9781139519526>

Vieyra Software. (2025). *Physics Toolbox*. <https://www.vieyrasoftware.net/>

Vieyra, R., Megowan-Romanowicz, C., O'Brien, D., Cortés, C., & Johnson-Glenberg, M. (2023). Harnessing the digital science education revolution: Smartphone sensors as teaching tools. In *Handbook of Research on Digital Teaching and Learning* (Chapter 8). IGI Global. <https://doi.org/10.4018/978-1-6684-5585-2.ch008>

Vieyra, R. E., Megowan-Romanowicz, C., Johnson-Glenberg, M. C., O'Brien, D., & Cortés, C. V. (2024). Making motion meaningful: Mapping body movements onto graphs. *The Science Teacher*, 91(6), 57–64. <https://doi.org/10.1080/00368555.2024.2404956>



Impact of Varying the Playback Speeds of Educational Content on Learning and Engagement

Tyree Cowell, Aaron Wong, Caitlin Mills, April Murphy, Stephen Fancsali, Rae Bastoni, & Steve Ritter

Eye-tracking

Mathematics Education

Online Learning

video engagement

As digital learning tools and pre-recorded lectures become more pervasive in a post-pandemic world, understanding the impact of playback speeds on learning has become more important. MATHstream, a video-based digital learning tool from Carnegie Learning, leverages the power of instructional videos to improve math outcomes. Building on previous literature and feedback from users, our team developed an experiment to measure the impact of different video speeds on engagement, enjoyment, and learning. The experiment conducted on 159 adults showed that there was no significant impact on eye-tracking measures, self-reported engagement measures, or performance on related math problems across video speeds. However, one instructor's videos were nearly twice as long as the others, leading to lower engagement, increased boredom, and lower likeability, suggesting that the length of a video has a greater impact on engagement than the speed at which it is played.

Tyree Cowell¹, Aaron Wong², Caitlin Mills³, April Murphy⁴, Stephen Fancsali⁵, Rae Bastoni⁶, Steve Ritter⁷

¹ Carnegie Learning; tcowell@carnegielearning.com

² University of Minnesota; wonga@umn.edu

³ University of Minnesota; cmills@umn.edu

⁴ Carnegie Learning; amurphy@carnegielearning.com

⁵ Carnegie Learning; sfancasli@carnegielearning.com

⁶ Carnegie Learning; rbastoni@carnegielearning.com

⁷ Carnegie Learning; sritter@carnegielearning.com

Abstract.

Keywords: online learning, math education, video engagement, eye-tracking

Introduction

MATHstream is an innovative digital learning tool from Carnegie Learning, a leading provider of K-12 EdTech solutions used by over 50,000 students in the 24/25 school year. MATHstream provides supplemental math instruction that uses videos, interactive questions, and gamification to support math learning. Feedback from students and teachers during site visits to research partner districts suggested that content in some videos was presented too quickly, especially for English second-language learners, and that having a way to slow playback speed would be helpful. Conversely, previous research on the impact of increased video speeds on learning has actually found that increased video speed (1.25x - 1.5x) is optimal for learning, particularly for young adults (Lang et al, 2020; Mo et al, 2022; Murphy et al, 2023; Liu & Jia, 2025). In response to these conflicting insights, our team sought to investigate the impact of different playback speeds on performance and engagement in the context of MATHstream.

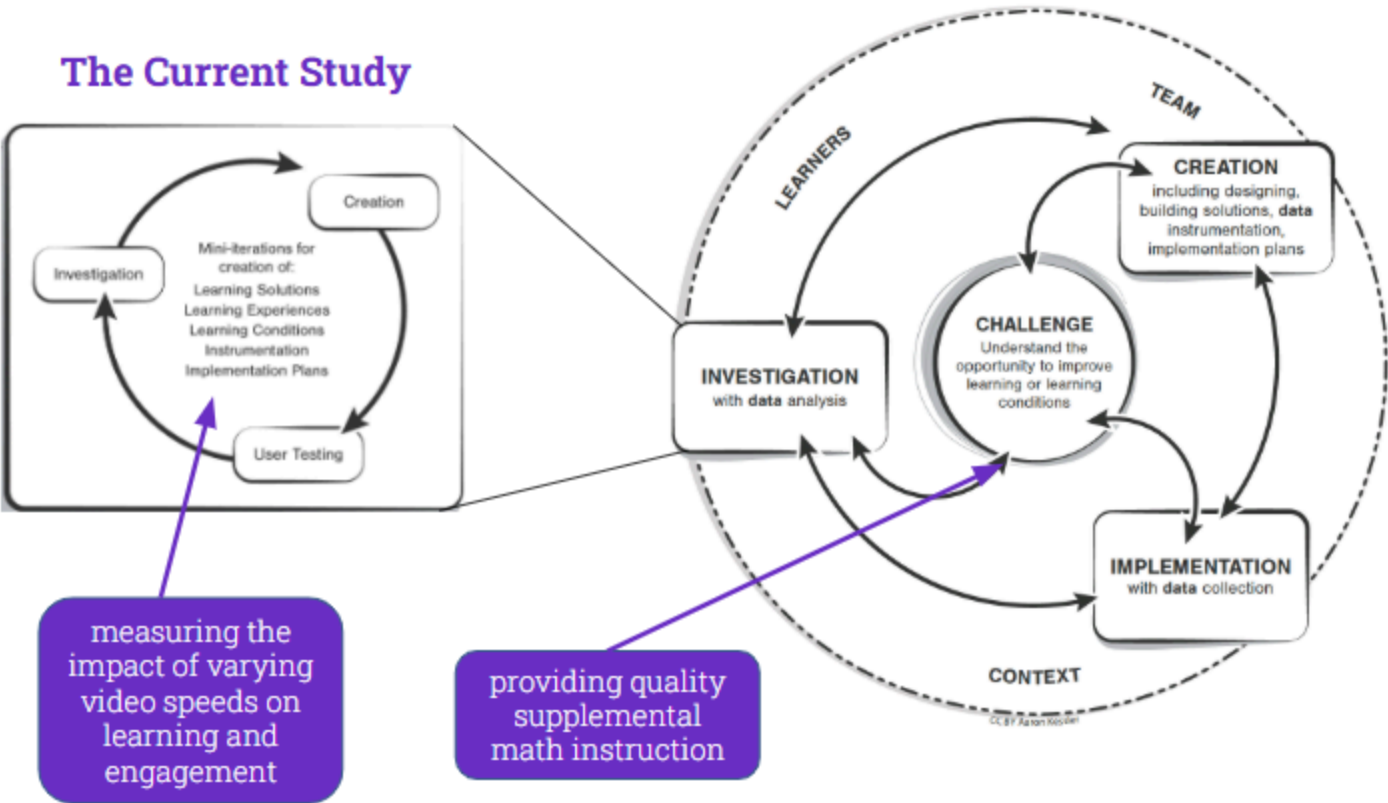
While MATHstream's target audience is primarily middle school and high school students, we implemented a pilot experiment with adults on the online testing platform called Prolific to get preliminary insight into whether varying playback speeds on the platform was a viable option for improving students' learning experience before making more substantial changes to the MATHstream platform. In the current study, we exposed participants to videos from MATHstream at 3 different speeds (slow, normal, and fast) in order to measure any impact the changes in speed might have on engagement and math performance.

Learning Engineering Process

The figure below presents the Learning Engineering Process as laid out by Kessler et. al. (2023). The process involves multiple, iterative phases moving through identifying challenges, creating solutions, implementing solutions, and investigating impact. Recent applications of this process have since identified that nested loops of this process may exist at various points during the cycle (Craig et. al., 2025). In the context of this study, the larger cycle represents MATHstream as a product solution where the challenge is providing quality supplemental math instruction in ever-changing classroom environments. During the investigation phase of that outer loop, we are able to use data we gathered from both the platform and user feedback to identify new, smaller challenges. In this case, the new challenge is that some students desire slower video speeds, while some research suggests that faster video speeds are optimal for learning. In order to address this challenge, we enter into a nested loop of creation, experimentation, and investigation.

Figure 1. Placement of current research in the Learning Engineering Process

MATHstream Product Solution



Methods

Participants

The study was conducted with 159 adults who were non-MATHstream users, and who were recruited through the online testing platform Prolific.

Materials and Procedure

Each participant was randomly assigned to one of nine experimental conditions described in Table 1. Participants viewed a short instructional video on Least Common Multiples taught by one of three instructors and presented in one of three playback speeds (0.75x, 1x, 1.25x). Instructional content was consistent across videos, though the exact scripts varied slightly by instructor.

Table 1.

All 9 conditions combining unique instructors and playback speeds.

Instructor A	Instructor B	Instructor C

0.75x (slow) speed	Instructor A Slowed Down	Instructor B Slowed Down	Instructor C Slowed Down
1x (normal) speed	Instructor A Control	Instructor B Control	Instructor C Control
1.25x (fast) speed	Instructor A Sped Up	Instructor B Sped Up	Instructor C Sped Up

Building on prior work from the University of Minnesota (Wong et al, 2023; Hutt et al, 2024), our procedure used webcams to capture eye-tracking data across 3 metrics: Area of Interest Proportion (AOIProp) refers to the proportion of gazes on the video, off-screen proportion (OffscreenProp) refers to the proportion of gazes not looking at the screen, and unique proportion (UniqueProp) refers to the proportion of gazes that are on unique locations. During the video, participants were probed for engagement every 60-80s. Participants were shown 4 questions in order to measure Task Unrelated Thought (TUT) (whether they were thinking about the video or something else), Difficulty Disengaging (would it be difficult to disengage from their current thoughts), Valence (how are they feeling on a scale from 1 = very negative to 7 = very positive), and Boredom (how bored they feel on a scale from 1 = very bored to 7 = not at all bored).

After watching the video, participants were asked to complete a short 15-question math quiz on the math content covered in the video (Least Common Multiples). Finally, they were asked to fill out a short survey to gather feedback on four usability metrics: Difficulty (rate the difficulty of the math problems on a scale of 1-5), Understandability (rate how well they could understand the instructor on a scale of 1-5), Likeability (rate how likeable they found the instructor on a scale of 1-5), and Desire for Speed Change (whether they would slow down or speed up the video). We hypothesized that participants in the sped-up conditions would demonstrate higher disengagement and lower performance on related multiple choice questions when compared to participants in the slowed down condition. We also predicted that these effects would be more pronounced for instructor Instructor B, who was reported by users to speak particularly quickly.

Results

In general, participants that watched videos taught by Instructor A demonstrated and reported lower engagement than those who watched videos with Instructor B and Instructor C. However, this may be confounded with video length as videos taught by Instructor A were markedly longer than those taught by Instructor B or Instructor C (average length of 634 seconds compared with 235 and 333 seconds, respectively). Video playback speed did not have a significant effect on most measures suggesting that 1x speed is sufficient for both engagement and performance. Finally, in a methodological sense, this study provides initial evidence that gaze behavior while watching videos can be used to predict self-reported engagement.

Eye-Tracking Measures

As shown in Table 2, there were no significant differences on any eye tracking metric across speed conditions. However, comparing different instructors, we saw that there was a significantly higher proportion of off-screen gazes for Instructor A compared to Instructor B and Instructor C ($p = 0.02$). Conversely, there was a significantly lower proportion of on-screen gazes for Instructor A compared to Instructor B and Instructor C ($p < 0.01$, $p = 0.03$).

Table 2.

Average values and standard deviation for eye-tracking measures across conditions.

Instructor	Speed	AOIProp	OffScreenProp	UniqueProp
DK	0.75	0.58(0.21)	0.14(0.13)	0.46(0.07)
DK	1	0.54(0.25)	0.21(0.25)	0.49(0.04)
DK	1.25	0.51(0.24)	0.16(0.11)	0.49(0.15)
HH	0.75	0.59(0.26)	0.19(0.20)	0.51(0.13)
HH	1	0.57(0.21)	0.17(0.16)	0.47(0.04)
HH	1.25	0.60(0.21)	0.13(0.12)	0.53(0.20)
RA	0.75	0.34(0.16)	0.35(0.19)	0.50(0.14)
RA	1	0.48(0.12)	0.19(0.13)	0.46(0.17)
RA	1.25	0.49(0.22)	0.24(0.22)	0.46(0.07)

Self-Reported Engagement Measures

Similarly to the eye-tracking measures, there were no statistically significant differences in self-reported engagement measures across different speed conditions. However, we observed a significantly higher proportion of self-reported task-unrelated thought (TUT) for participants who watched Instructor A's videos compared to videos taught by Instructor B and Instructor C ($p = 0.019$, $p < 0.01$) (Table 3). Additionally, users were more likely to report positive feelings towards Instructor C compared to Instructor A ($p < 0.01$) and were more likely to report feelings of boredom towards Instructor A compared to Instructor C ($p < 0.01$).

Table 3.

Mean self-reported engagement measures and standard deviation across conditions.

Instructor	Speed	n	Videolength(s)	TUT	Disengage	Valence	Boredom
DK	0.75	21	301	0.19(0.19)	0.21(0.32)	5.13(1.30)	4.65(1.83)
DK	1	20	225	0.15(0.29)	0.13(0.28)	5.08(1.03)	4.68(1.65)
DK	1.25	17	180	0.24(0.36)	0.06(0.17)	4.94(1.29)	4.41(1.5)
HH	0.75	16	425	0.16(0.13)	0.13(0.18)	5.43(1.17)	5.16(1.34)
HH	1	20	319	0.06(0.14)	0.16(0.33)	5.5(1.060)	4.95(1.84)
HH	1.25	15	255	0.13(0.21)	0.29(0.38)	5.47(1.29)	5.49(1.48)
RA	0.75	16	810	0.33(0.28)	0.13(0.24)	5.04(1.25)	3.93(1.68)
RA	1	17	607	0.36(0.28)	0.04(0.08)	4.50(1.38)	3.67(1.81)
RA	1.25	17	486	0.26(0.20)	0.09(0.18)	4.62(1.42)	4.21(1.84)

Math Performance and Video Feedback

The results from participants' scores on the 15-question post-quiz on Least Common Multiples and post-video survey responses indicated no significant differences between different playback speeds (Table 4). We noted that at normal playback speed, participants rated Instructor A as less likeable than Instructor B or Instructor C ($p < 0.045$)

Table 4.

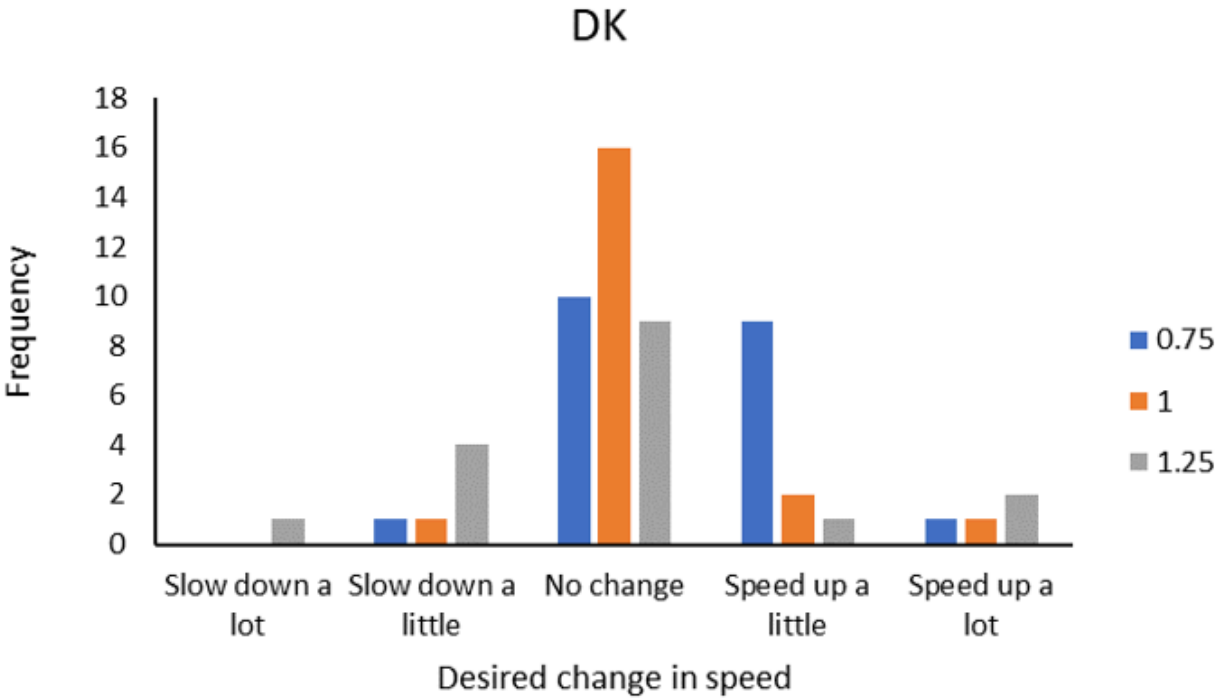
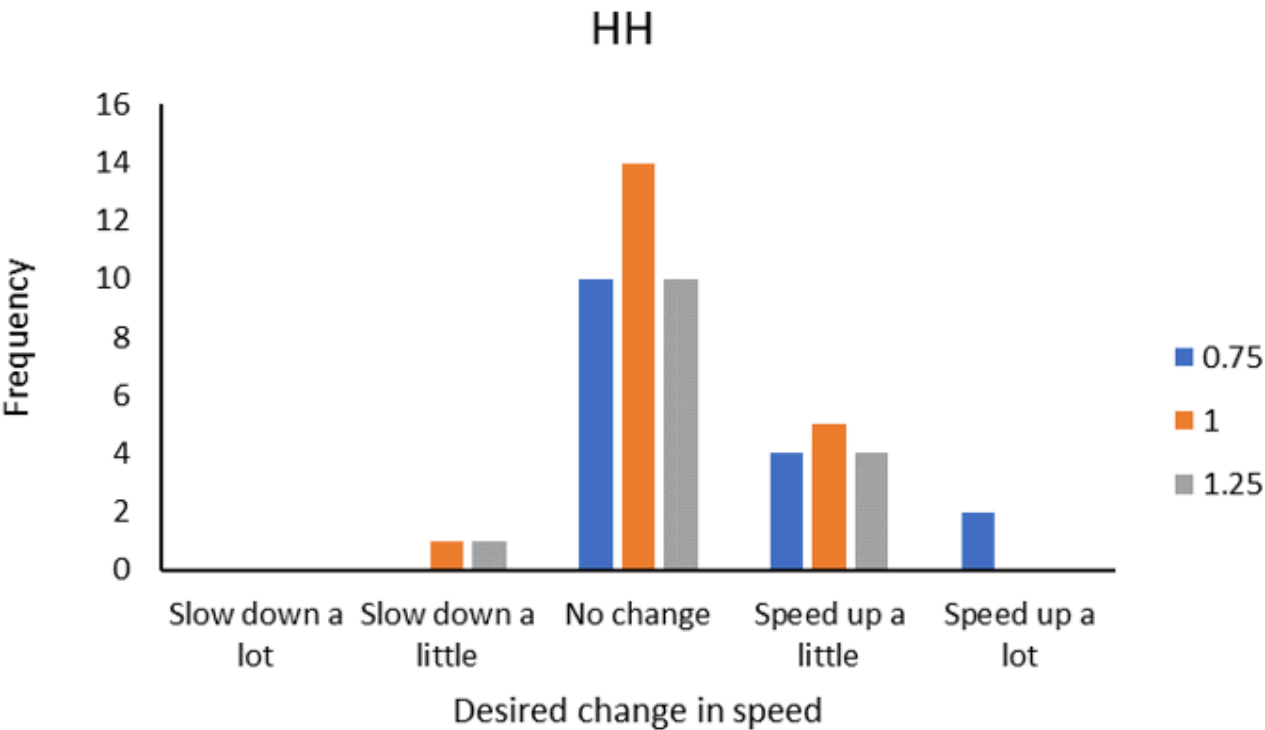
Mean quiz performance and survey feedback with standard deviations across conditions

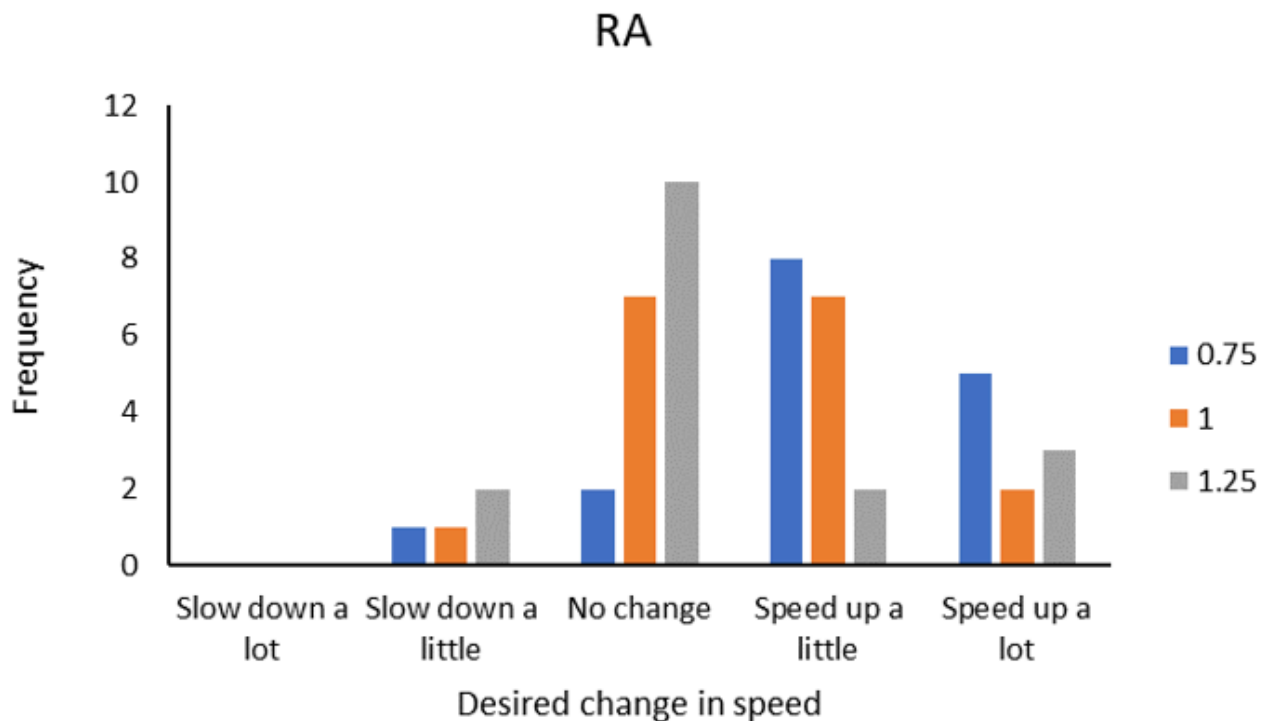
Instructor	Speed	MQScore	Difficulty	UnderstandEase	Likeability	SpeedChange
DK	0.75	12.76(2.19)	2.48(1.12)	4.43(0.75)	4.67(0.48)	3.48(0.68)
DK	1	12.35(2.78)	2.25(1.16)	4.50(0.76)	4.70(0.47)	3.15(0.59)
DK	1.25	13.47(1.74)	1.94(0.97)	4.18(1.01)	4.35(0.79)	2.94(1.02)
HH	0.75	13.50(1.59)	2.44(1.09)	4.50(0.82)	4.19(1.17)	3.50(0.73)
HH	1	12.90(2.25)	2.25(1.21)	4.65(0.59)	4.70(0.47)	3.20(0.52)
HH	1.25	11.80(2.73)	2.20(1.15)	4.27(0.88)	4.60(0.51)	3.20(0.56)
RA	0.75	13.69(2.52)	2.00(0.89)	4.63(0.62)	4.31(0.79)	4.06(0.85)
RA	1	12.29(3.44)	2.65(1.27)	4.41(1.00)	4.00(0.61)	3.59(0.80)
RA	1.25	12.35(3.33)	3.00(1.41)	4.12(0.99)	4.59(0.51)	3.35(0.93)

Desire for Speed Change

Finally, we asked participants in each condition if they would prefer the video to be faster, slower, or keep it as is (no change). For instructors Instructor C and Instructor B, most participants were satisfied with the speed to which they were assigned. While we expected that participants would prefer Instructor B's instruction in the slow condition, surprisingly, we found that they often wanted her to speed up. Finally, for instructor Instructor A, most participants preferred the sped up (1.25x) speed and participants in Instructor A's normal speed (1x) condition wanted him to speed up.

Figure 2. Results for survey question asking participants if they wanted the video they watched to be sped up, slowed down, or no change.





Discussion

Limitations

The most conspicuous limitation of this study is that it was conducted with adults, whereas users of the MATHstream platform are primarily middle school students between the ages of 11-14. Additionally, participants did not watch videos in the larger context of using the MATHstream platform in a classroom environment as part of either core or supplemental instruction. We have since conducted follow-up studies with middle school students who are MATHstream users, with which we compared these results.

Conclusions

This study showed that changing the video playback speeds from 1x to 0.75x or 1.25x did not have significant impacts on user engagement via eye-tracking or self-reported engagement measures, or on math performance. Most differences in engagement were found when comparing measures across different instructors. Specifically, participants who watched videos taught by Instructor A showed lower on-screen gaze behavior, higher reports of boredom, and lower likeability scores. A likely explanation for these results is that videos taught by Instructor A were substantially longer than those taught by Instructor B and Instructor C. This suggests that the length of the instructional video has a larger impact on participant engagement than the speed at which the instructor is speaking.

In most cases across the conditions, participants reported having no desire to change the speed of the video. These results indicated that the existing 1x speed was sufficient for both learning and engagement in this context. We found this result to be somewhat surprising as several studies (Lang et al, 2020; Mo et al, 2022; Murphy et al 2023; Liu & Jia, 2025) have found 1.25x and 1.5x speed to be more optimal for learning than 1x speed. An additional study from this team, conducted directly in the MATHstream platform with middle school users replicating the 3 speed conditions, found no significant differences in student

performance or content completion rates (Cowell et al., 2024). These findings provide evidence that user engagement and performance when watching instructional videos may not be as sensitive to playback speed as other work has suggested (Lang et al, 2020; Mo et al, 2022; Murphy et al 2023; Liu & Jia, 2025). Instead, factors such as instructor quality and likeability may be more likely to drive these outcomes. This has implications in designing video-based educational materials, particularly when course lessons involve multiple instructors delivering lesson content.

References

- Cowell, T., Mehta, S., Fancsali, S., Ritter, S. (2024). Variant Playback Speeds in MATHstream: An In-Platform A/B Test. [Unpublished Report].
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). <https://doi.org/10.59668/2109.21735>
- Hutt, S., Wong, A., Papoutsaki, A., Baker, R. S., Gold, J. I., & Mills, C. (2024). Webcam-based eye tracking to detect mind wandering and comprehension errors. *Behavior Research Methods*, 56(1), 1-17.
- Kessler, A., Craig, S. D., Goodell, J., Kurzweil, D., & Greenwald, S. W. (2023). Learning engineering is a process. In *Learning Engineering Toolkit*. Taylor & Francis.
- Lang, D., Chen, G., Mirzaei, K., & Paepcke, A. (2020, March). Is faster better? A study of video playback speed. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 260-269).
- Liu, R., & Jia, H. (2025). Effects of instructional video playback speed and pre-embedded questions on learning: R. Liu and H. Jia. *Asia Pacific Education Review*, 26(2), 383-395.
- Mo, C. Y., Wang, C., Dai, J., & Jin, P. (2022). Video playback speed influence on learning effect from the perspective of personalized adaptive learning: A study based on cognitive load theory. *Frontiers in Psychology*, 13, 839982.
- Murphy, D. H., Hoover, K. M., & Castel, A. D. (2023). The effect of video playback speed on learning and mind-wandering in younger and older adults. *Memory*, 31(6), 802-817.
- Wong, A. Y., Bryck, R. L., Baker, R. S., Hutt, S., & Mills, C. (2023, March). Using a webcam based eye-tracker to understand students' thought patterns and reading behaviors in neurodivergent classrooms. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 453-463).

