Advancing Research on Equitable AI Education Through a Focus on Implementation: Insights from a Middle School Computer Vision Module Beta-Test

Christina A. Bosch, Mary Cate Gustafson-Quiett, Samar Abu Hegly, Sarah Wharton, John Masla, Lydia Guterman, Calvin Macatantan, Eric Klopfer, Hal Abelson, Cynthia Breazeal

Massachusetts Institute of Technology

cabosch@mit.edu, marycate@mit.edu, samarh18@mit.edu, wharton@mit.edu, j_masla@mit.edu, lyd@mit.edu, cmacatan@mit.edu, klopfer@mit.edu, breazeal@mit.edu, hal@mit.edu

Abstract

Part of a university initiative supporting responsible AI for social empowerment and education, the project-based RAICA (Responsible AI for Computational Action) curriculum supports middle/high school learners and novice AI literacy teachers use AI creatively for good. This paper offers a rare example of design-based implementation research (DBIR) in AI education across widely varied contexts, provides fine grain implementation data that contributes to a foundation for evaluating effectiveness and expanding access. We present a novel approach to analyzing fidelity of implementation data from RAICA's computer vision module beta-test. Twelve educators working with ~282 students across nine pilot sites in four countries used a bespoke fidelity of implementation data collection tool (pre-made comment prompts in a Google Docs version of the teacher guide) to provide 236 qualitative responses about AI literacy and responsible design activities, plus 111 ordinal ratings of embedded teacher supports. Analyses revealed that while the curriculum was generally implemented as designed, educators frequently made modifications. Although most changes produced practical insights for improved curriculum design, others helped the design team anticipate and prevent changes that could obscure learning objectives and hinder outcomes. We discuss the pedagogical, design, and research implications of these findings for effective AI teaching/learning in diverse settings.

Introduction

Implementation research helps explain not only whether an intervention works, but also why it works, for whom, and under what conditions—questions that are increasingly important as researchers aim to iteratively design and understand nascent AI education efforts, especially with populations traditionally marginalized in STEM fields and AI industry. Although different models of implementation research exist in education, a shared focus is on capturing data about both context and curricular components (e.g., Hill et al., 2012; AIR, n.d.; Allison, 2023), which will be key to generating conclusions that address tensions and calls to action raised in literature on teaching AI to K-12 learners (e.g., Grover, 2024).

Implementation research is a critical frontier for AI education studies interested in impacting teacher practice, uncovering effective or unique aspects of relevant pedagogies, and understanding the learning trajectories or conceptualizations involved in learning about AI across developmental stages. Consider, for instance, that recent attention to implementation in AI education has yielded the important finding that most AI literacy in K-12 between 2020 and 2022 focused on technical education rather than application of such knowledge to create or evaluate (Olari et al., 2023). In a subsequent review of AI literacy frameworks, Almatrafi et al. (2024) showed a shift in the field that responded to those disparities in instructional aims, while also revealing new ones. The authors found that "recognize" as well as "know and understand" (which most curriculum developers would consider foundational constructivist learning objectives) figured less frequently than higher-order skills like "use and apply." "Create" was even less common, a dissonant result given that constructionist learning theories or project-based pedagogies are often invoked to explain the design of AI education interventions (Almatrafi et al., 2024). Thus, the current paper contributes a model of design-based implementation research (DBIR) to inform not only the quality of AI curricula but ultimately the understanding of what, if anything, is novel and/or unique about teaching and learning AI in K-12.

Reviews of current AI literacy initiatives in K-12 education show that more attention to implementation could expand and improve practical, pedagogical understandings of AI education. In their 2022 paper, Yue et al. found that most AI literacy studies did not focus on implementation with teachers in educational settings, were less than a week long in total, and were conducted outside of the regular school day. Almatrafi et al. (2024) noted in their systematic review that about half of the K-12 papers lacked empirical evaluation of effectiveness. While important data on student demographics, prior knowledge, and instructional conditions were reported by the Su & Yang (2023) paper included in the review, it was not consistent in the others. Studies such as Buxton et al. (2024) and Radday & Mervis (2024) have shared rich curricular details, but without detailed empirical insights into how these interventions were taken up in different educational contexts, even in exploratory research stages.

The current study introduces a novel approach to implementation research in the RAICA (Responsible AI for Computational Action) project, exploring how design-based implementation research methods can support goals related to access, equity, and quality in AI education initiatives, which can strengthen the evidence base linking AI curriculum interventions to educational outcomes and pedagogical insights. We discuss the implications for curriculum design, teachers/learners, and the broader AI K-12 community through the following questions:

1. Which components of the curriculum were implemented?
2. How did implementation vary within the curriculum and across contexts?
3. How did collaborating educators respond to the components of the curriculum?

Curricular and Conceptual Components of RAICA

AI literacy frameworks seem to have coalesced around the importance of higher-order skills like using, evaluating, and/or creating with AI (Ng et al., 2021; Almatrafi et al., 2024). Congruently, the RAICA curriculum contributes to developing "AI fluency," a long-term goal involving conscientious engagement with and ethical design of AI, as well as participation in democratic discussions around artificial intelligence.

RAICA is a project-based curriculum that includes facilitator guides, classroom slides, and resources that help novice AI students and teachers learn how some AI technologies work. Students and teachers only need introductory experience with block-based coding, like Scratch, to engage with scaffolded opportunities to apply computational thinking concepts, skills and practices (Brennan & Resnick, 2013) through content lessons and supported "responsible design" of original projects that focus on stakeholders, values, and impact (Wharton et al., 2024). Students make self-determined projects in a Scratch-like coding platform with AI extensions – the RAISE Playground (MIT, 2024). Supports for technological (T), pedagogical (P), and content (C) knowledge (TPACK) development are inserted at key opportunities for developing the seven intersecting forms of teacher knowledge (i.e., T, P, C, TP, TC, PC, TPCK; Mishra et al., 2023) and/or to support student and teacher co-learning (Brantmeier, 2005; Crompton, 2023).

This constructionist (Papert, 1980) approach drives students beyond thinking and creating towards computational action based in AI literacy, to explicitly "make computing education more inclusive, motivating, and empowering for young learners" (Tissenbaum et al., 2019, p. 34). Towards that end, RAICA is, to our knowledge, the only AI literacy curriculum that employs Universal Design for Learning (UDL)[1] as a framework to proactively design and user-test for accessibility, cross-cultural contexts, and a wide range of learning environments.

RAICA's suite of materials (i.e., slides, teacher guide, graphic organizers and scaffolds, computational tools) form an integrated curricular activity system to meet teacher, student, and diverse learning environment needs (Vahey et al., 2013). The curriculum is comprised of modules that each focus on topics like natural language processing, reinforcement learning, and in the case of the current study, computer vision and image classification. Modules are divided into parts intended to take 12 – 15 hours of instructional time total. Each part consists of three core curriculum components: activities, formative assessments, and – in the teacher guide – embedded "teacher notes" that may be short textual reminders or links to additional supportive resources. These three types of curricular components address AI literacy, responsible design, computational action, and TPACK concepts. Effective design and implementation of these four broad concepts is critical to achieving the curriculum's corresponding outcomes.

While all four major concepts are woven throughout all eight parts of the RAICA "Picture This" module, AI literacy is the focus of the first four parts, and responsible design as well as computational action are the focus of the last four, with associated aspects of TPACK emphasized accordingly. Responsible design is the only concept that appears as an explicit label of the content in the curricular materials. The number of items within each part in Table 1 refers to the required and optional items in the fidelity tracker (FT) instrument described in the Methods section, but because FT

[1] https://udlguidelines.cast.org/

Part No.	Part Name (and topics)	No. FT Items
1	Launch	20
2	Image Classification (definitions, models, data bias)	26
3	Computer Perception (RGB values, pixels, human vs computer image processing)	22
4	Neural Networks (model training)	22
5	Start your Projects (stakeholder, design values, & impact ideation)	16
6	Prototype	16
7	Try it Out	16
8	Showcase (stakeholder, design values, & impact identification)	12

Table 1. Overview of content available for implementation in RAICA module

items align with curricular components, it gives a sense of the amount of content in each part.

The teacher guide provides a curriculum overview along with standards-aligned daily lessons that explicated preparation, materials (e.g., slides, rubrics, graphic organizers), procedures, and additional tips.

Methods

Design-Based Implementation Research (DBIR) is a systematic approach that engages stakeholders in iterative research to investigate and improve the implementation process, ensuring it addresses local needs and challenges (Fishman et al., 2013). Unlike methods focused solely on testing a specific intervention, DBIR explores how to design and adapt implementation processes that respond to multiple challenges within the complexity of education systems and settings. This flexibility makes it well-suited for research on innovative approaches in AI education, where multiple barriers to acquiring content and technical knowledge and skills can combine with contextual factors such as varying levels of teacher expertise and school resources to significantly impact outcomes.

DBIR emerged from Design-Based Research (DBR), which was originally developed by learning scientists to test educational interventions in real-world settings (Christensen & West, 2018). While DBR focuses on understanding how specific solutions address educational challenges, "researchers engaged in DBIR not only investigate and build theory about an innovation's impact on learning, but also about how and why an innovation is implemented differently in different settings" (Fishman et al., 2013). While the RAICA project is not conceived of exclusively as a DBIR project due to limitations in capacity, DBIR is particularly relevant for the current stage of the project as it seeks to overcome problems of practice that tend to persist in project-based learning, engage in collaborative, iterative design in authentic, diverse educational settings, build theory and knowledge around both learning and implementation of AI interventions, as well as developing capacity locally and remotely for sustaining AI teaching/learning.

Implementation Data Collection

Given the lack of available AI education implementation tools and the unique emphases on responsible design and computational action in RAICA, the project team developed a custom "Fidelity Tracker" (FT) tool based on a multi-stakeholder model of fidelity of implementation (Hemmeter et al., 1996; Swain et al., 2013). Implementation questions about how curricular components were used were embedded into the teacher guide (a cloud-based document) as comments with multiple-choice and/or open-response options. Teachers typed in letter options (for multiple choice) or gave qualitative feedback. We provided a section on best practices for FT use, which included three bulleted guidelines describing the purpose of the Tracker as well as when and how to refer to the teacher guide and respond to the Tracker questions. Questions for each part asked for the date of implementation, number of students in attendance, selecting a multiple-choice option about the implementation of curricular components (identified as 'required'), and an optional explanation. The rationale behind the questions was to capture both fidelity and adaptations to the curriculum. Multiple-choice questions for the curricular components asked:

- How did you use this assessment/activity? (Options: Prescribed, Modified, Replaced, Skipped), to understand how closely the curriculum was followed and what adaptations were made.
- Teacher Notes: To what extent did the note support your understanding? (1 = Disagree, 3 = Agree, N/A), to gauge the utilization of provided instructional guidance.

Teachers were encouraged to provide explanations for their responses, allowing us to better understand the context of any modifications. To ensure regular data collection, we held meetings with a subset of teachers and maintained asynchronous communication, prompting responses if the tracker was not completed within two weeks.

Sample

A maximal variation sampling strategy was used to identify and collaborate with a wide range of educational settings (i.e., geographic location, funding, instructional context) to support a flexible, translatable design for the curriculum.

Participating educators were recruited through a range of incoming and outgoing communication channels and existing networks. If initial email exchanges showed that potential participants taught computer science or digital literacy directly to students (including students with identified disabilities) in English or Spanish at the middle school level, they were invited to an interview with two team members to discuss alignment between participant's instructional goals and practical needs with the project's characteristics. Candidates from fourteen potential sites were interviewed and nine sites retained. Three sites often had two to four teachers involved in the class (totaling 12 participants considered as teachers), and did not have a 1:1 device ratio.

The sites across four countries (U.S., Malawi, Chile, and UAE), covering grades 6-12, included private, parochial, public and charter schools; an individualized special education setting; and a nonprofit community development center in a refugee camp. Class types varied from required courses (e.g., Computer Science, Technology, Digital Literacy, and Math) to electives, with class durations ranging from 45 minutes to 1.3 hours and frequencies of once a week to daily sessions. Not counting the individualized instruction participant, enrollment across sites ranged from 10 to 67 students per class (totaling approx. 282 students), with special education representation between 7% and 58%. Teachers varied in gender and racial demographics, age and experience levels, with a mix of new and experienced educators, and diverse backgrounds including subject matter expertise in art, technology and robotics. Four had taught about AI before.

Analysis

Analysis was conducted by the RAICA research team, consisting of five core members with expertise in curriculum design and research, working with an undergraduate and a graduate researcher assistant on data management.

Data from each teacher's fidelity tracker, which included both ordinal (i.e., multiple-choice selections) and qualitative data (i.e., responses to comments), were consolidated into a database. Data cleaning included a team-based review, comparing the original data collected with the database entries. A subset of the team calculated completion rates through pivot table analyses of ordinal data. Completion rates were calculated as (# completed ordinal responses) / (# completed multiple choice responses + # no answer ordinal responses). We excluded any responses from teachers who did not do a given part of the module. Similarly, no rate was calculated for comments because they were optional.

For the qualitative data, an inductive coding scheme was initially developed by the researcher for the qualitative data, designed to quickly capture the concepts expressed in teacher comments to distinguish information relevant for curricular evaluation and revision. The team then calibrated the coding scheme through a collaborative review process, where multiple team members coded the same sample of data and compared results to ensure alignment and consistency in applying the codes. Comments related to activities, assessments, and teacher support were reviewed and coded by the curriculum designer, assessment specialist, and teacher support specialist, respectively. Each comment received an initial code, and further value codes were applied to those labeled as "teacher adaptation" to indicate whether the adaptation may have had a positive, negative, or neutral impact on the learning objectives.

Results

Five out of nine sites completed all eight parts of the module. Every site made it through the first three parts, one dropped off after part three, two more schools dropped off after part four, and one more after part six.

Teachers responded to the multiple-choice questions 77.78% of the time and stayed above a 70% response rate for each part except part 8, the last part, which had a multiple-choice response rate of 46.15%. The FT captured 225 qualitative responses from participants across the entire module. Part 2 generated the highest number of comments (54), and Part 8 generated the lowest (7).

Teachers responded to multiple choice questions about activities, assessments, and teacher notes at similar rates (77% +/-2). The 32 activities throughout the module generated the most qualitative responses (125), followed by the 22 teacher notes (55), and the 11 assessments (47).

Variation in Implementation: U.S. & International

In a preliminary analysis of how implementation varied between the five U.S. and four international settings, we found the former skipped activities more frequently (17 vs 3% of the time), modified collaborative student groupings less frequently (2 vs 5% of the time), and implemented with slightly less fidelity (37 vs 47% of the time). Similarly, U.S. schools implemented assessments with much less fidelity (27% vs 50%) even though modifications to assessments were somewhat comparable (36 vs 31%). Of the parts of the module completed by each participating site, U.S. sites skipped assessments 18% of the time while international schools never did. U.S. schools gave more "neutral" ratings to teacher notes (15% vs 7%) only international sites ever disagreed with their utility.

Fidelity and Feedback by Curricular Component

Fidelity by curricular component examined how each element of the curriculum—such as activities, assessments, and teacher notes—was implemented, based on teachers' multiple-choice responses. These responses indicated whether components were implemented "as prescribed" or modified. This analysis provided the team with valuable insights into

which components required revision for improvement and which could be refined or expanded based on positive teacher feedback. Teacher comments further pinpointed specific, actionable areas for enhancement, including curriculum examples, instructions, vocabulary support, technical troubleshooting, and increasing easy of navigation between the curricular activity system materials.

Activities

Participants reported implementing RAICA activities "as prescribed" 40.7% of the time. The most common modifications reported were "modified activity" (24.9%) and "modified timing" (17.5%). Fidelity rates varied across sites, with the highest reported fidelity at 51.85% and the lowest at 20%. The activity implemented with the highest fidelity was Activity 3.4.0, an introduction to computer perception, where six out of seven teachers implemented it "as prescribed," and one teacher reported "modified timing." Activities 5.5.0 (the "define" step of responsible design) and 6.2.0 (intro to prototyping) were consistently implemented "as prescribed" in all five instances. In contrast, no participants reported implementing activities 4.5.0 (acting out a neural network) or 8.3.0 (project showcase) "as prescribed."

During beta-testing, 119 comments were collected from nine teachers about the prescribed activities. Initial codes were applied to summarize the content of these comments. The most frequent code, "teacher adaptation [we should review and evaluate]," appeared 42 times. Thirteen activities received 15 comments indicating the need for improved instructions, examples, links, or vocabulary. Additionally, five instances of student work being shared were noted. Activity 4.4.0 (intro to neural networks) received the most comments (n=8), including "teacher adaptation" (n=3) and "improve instructions" (n=2), while Activity 2.3.0 (intro to image classification) followed with seven comments (n=7), mostly related to "teacher adaptation" (n=3) and "accessibility" (n=2).

Over half of the adaptations were value-coded as neutral, nine out of 42 were found to be positive adaptations that supported the learning objective, and six adaptations were coded as having a negative impact on the stated learning objectives.

Assessments

Participants reported implementing RAICA assessments 'as prescribed' 37.5% of the time. 'Modified content/format' was the second most common teacher response. Of the four teachers who responded to Parts 5 and 6 exit tickets, none reported implementing 'as prescribed,' and only one out of eight teachers reported implementing the Part 2 exit ticket 'as prescribed' while six reported modifications. 18 out of 24 'teacher adaptations' were value coded as 'neutral,' three were found to be positive, and two 'negative.' Twenty out of 36 total teacher comments on assessments were coded as teacher adaptations. The Part 2 exit ticket generated the most comments and all seven comments included insights into teacher adaptations.

Teacher Notes

73.9% of the time teachers agreed that the teacher notes were supportive to their understanding of the content or the tools. While the introductory image classification lesson teacher note (3.1.0) was most the most frequently rated as helpful, the immediately subsequent note (3.2.0) along with a support for prototyping (6.4.0) received the only three 'disagree' ratings.

The most used code for teacher comments was "supported teacher understanding" (30.6%). Teachers also shared six adaptations across six different teacher notes. Additionally, there were 14 instances where teachers highlighted specific areas in the curriculum needing improvement, such as instructions, examples, vocabulary clarification, and technology troubleshooting support.

Fidelity and Feedback by Curricular Concept

The curriculum's learning objectives stem from three constructs that describe intended outcomes: AI literacy, computational action, and responsible design. When items were analyzed across parts and grouped according to conceptual components, the fidelity tracker showed that items about AI literacy and responsible design components were completed 81% of the time, while those about computational action were completed 69%. However, the number of items available to respondents for AI literacy (n=162) greatly exceeded those about responsible design (n=60) or computational action (n=37).

Nevertheless, computational action and AI literacy components were implemented as intended 19 and 24% of the time, respectively, while no responsible design components were. In fact, responsible design components were modified 75% of the time, whereas AI literacy and responsible design components were modified 29 and 13% of the time, respectively. Eight percent of computational action items were rated as "skipped or replaced" in contrast with AI literacy and responsible design (4%).

In terms of teacher knowledge, the fidelity tracker data showed comparable rates of completion for supports related to TPACK and its seven knowledge domains (76% on average). Importantly, there were many more opportunities for participants to complete items about TPACK (n=41) and PC knowledge (n=29), followed by T, P, TP and TC knowledge, than about content knowledge (n=7). Accordingly, teachers reported using notes about TPACK the most (n=41), followed by pedagogical, technological and TP knowledge in nearly equal measure (n=12). At the same time, TP knowledge appeared as the area most rated as not used (n=20).

If teachers used the notes, they indicated agreement, disagreement, or neutrality with their helpfulness. TPACK and PC knowledge notes received the most "agree" ratings (n=36 and 23, respectively), while all other domains of knowledge were rated this way between 4 and 9 times. Neutral ratings were evenly distributed across knowledge types. Disagree ratings were rare, occurring only across technological, pedagogical, and TP components.

Discussion

This implementation study revealed several insights into the delivery and adoption of the RAICA computer vision module. Of the nine sites, five completed all eight parts of the module, while the remaining sites dropped off at various points—one after part three, two after part four, and one after part six. Teachers demonstrated a high response rate (77.78%) to multiple-choice questions throughout the module, maintaining above 70% for most parts except part eight, where the response rate fell to 46.15%.

Qualitative feedback was robust, with 225 responses captured across the module. Part two generated the most comments (54), while part eight generated the fewest (7). Teachers responded to questions about activities, assessments, and teacher notes at similar rates (77% ±2%). Activities produced the most qualitative feedback (125 comments), followed by teacher notes (55) and assessments (47). Notably, fidelity rates varied between U.S. and international sites, with U.S. sites skipping activities and assessments more frequently.

Curriculum Implementation

Our first research question focused on identifying which components of the curriculum were implemented. Analysis of teacher fidelity trackers revealed that activities were implemented "as prescribed" 40.7% of the time, with common modifications involving changes to the activity itself (24.9%) and timing adjustments (17.5%). Fidelity rates varied across sites, ranging from 20% to 51.85%. Activity 3.4.0 (introduction to computer perception) had the highest fidelity, implemented by six out of seven teachers.

Two particularly substantive activities (4.5.0 and 8.3.0) were not implemented as prescribed by any site. The Part 4 embodied learning activity about neural networks was developed through multiple playtests with teachers and area experts. Given this background design work, the ubiquity of modifications to this activity suggests that it is one of the more complex in the module, which may reflect with the complexity of the subject matter. This particular AI topic seems to require that teachers invest in developing multiple domains of new knowledge. The modifications to the culminating project showcase (8.3.0) underscore a perennial challenge in executing project-based learning, which is the culminating presentation of projects to authentic audiences, ideally beyond the classroom (Blumenfeld et al., 1991). This challenge has implications for whether it is possible to achieve the community-impact tenet of computational action as an outcome of curricula implemented in structured group learning environments.

Assessments were implemented "as prescribed" 37.5% of the time, with modifications such as "modified content/format" being common. Despite having access to editable exit tickets in Google Docs, we found that teachers exhibited a preference for exit tickets as digital forms (e.g., Google Forms). None of the teachers implemented the exit tickets for parts five and six "as prescribed," and only one teacher did so for part two. Qualitative data showed that when were skipped, it was because teachers ran out of time at the end of class – not because they didn't find the assessments useful. These data raise questions about RAICA's assessment approach: providing teachers with ready-made formative assessments vs. supporting teachers to design their own (Grover, 2021).

Teacher notes were generally found useful (73.9% agreement rate) and seem to have been used to develop every domain of teacher knowledge in the TPACK framework. However, we realized that uneven coverage of supports for each domain in our materials coupled with a clear tendency to not use the technological-pedagogical (TP) supports signaled an area for future focus.

Variability Across Contexts

Regarding research question two, we observed that implementation differed not only between sites but also within various parts of the curriculum. U.S. sites skipped activities 17% of the time compared to 3% for international sites and implemented assessments with lower fidelity (27% vs. 50%). Additionally, collaborative activities or those requiring specific grouping structures were more frequently modified in international settings (5% vs. 2%). These differences may be due to differing cultural or logistical approaches to group work, and suggest a need for more inductive research into how assumptions and priorities about project-based AI education can vary internationally.

These discrepancies highlight potential contextual challenges, such as varying school structures, resource availability, and educator familiarity with AI concepts. For example, while international sites never skipped assessments, U.S. sites did so 18% of the time. Teachers in U.S. sites gave more "neutral" ratings to teacher notes (15% vs. 7%), while international sites occasionally disagreed with the utility of these notes.

As AI adoption grows globally, these findings underscore the need for curricula that can be widely adapted without

compromising core learning objectives. Developing universally designed resources (e.g., through UDL) can promote equitable access to AI education.

Educator Responses to Curriculum Components

Our third research question explored how collaborating educators responded to the curriculum. Teachers provided extensive qualitative feedback on the curriculum components. Only a few teacher notes (e.g., 3.2.0 and 6.4.0) were identified for improvement, whereas activities generated 125 comments, with "teacher adaptation" being the most frequent code (42 instances). Positive feedback highlighted the effectiveness of certain activities (e.g., 3.4.0), while other activities (e.g., 4.4.0 and 2.3.0) required clearer instructions and better examples. For assessments, 20 out of 36 teacher comments were adaptations, with most changes coded as neutral (18), three as positive, and two as negative. The part two exit ticket elicited the most feedback.

Clarifying instructions, enriching the examples provided, increasing vocabulary support, and guiding technology troubleshooting emerged as actionable areas for improvement. The prevalence of "teacher adaptation" codes regarding module activities indicates that teachers often found it necessary to modify the curriculum to meet their specific classroom needs. These insights could suggest a need for more robust scaffolding and differentiation options within the curriculum to better support diverse learning environments. Alternatively, they may highlight that modification and adaptation are a valuable pedagogical skill for teachers seeking to instruct students on AI, given that over half of adaptations were coded as “neutral” changes with no implications for learning objectives, and perhaps more importantly, that 22% of the adaptations represented helpful changes that advanced the established learning objective.

As discussions about how AI will affect teachers’ jobs proliferate, this kind of evidence may establish a foundation for future research on teacher pedagogy and praxis. We also found that in the case of modifications to assessments, teachers sometimes transformed them from individual opportunities for self-reflection to public processes (such as the raising hands to show responses as a whole-class, turning the exit ticket questions into a game or integrating them into the final presentation). These spontaneous changes align well with our emphasis on collaborative, project-based learning, which is exciting to see applied to the traditional individually oriented approach to formative assessment.

Limitations

While this study provides valuable insights into the implementation of the RAICA module, it is important to acknowledge its limitations. First, the data is limited to a self-reported measure of implementation fidelity, which is necessarily subject to bias and inconsistency. Second, while this experience report seeks to contribute a model of implementation research to the broader AI education community, the maximal variation sampling strategy limits any analysis into variability across types of educational settings beyond domestic (U.S.) and international. Third, reliability across coders was not established due to capacity limitations and the prioritization of our data-based revision process over empirical rigor. The coding scheme will also undergo future revisions to address inconsistencies in how the “pacing insight” code and associated value codes may have been applied.

Conclusions

Our experience in developing, implementing, and refining the RAICA curriculum provides some key contributions for AI education stakeholders. First, feedback collected during this phase of design-based implementation research (DBIR) informed revisions to the RAICA module, linking curricular components to target outcomes. For example, the data indicates that while AI literacy is achievable due to high completion rates of related activities, computational action may be compromised if concentrated near the end, where drop-off rates increase—a challenge also noted in project-based learning initiatives (Larmer & Mergendoller, 2011).

Second, feedback revealed consistent areas for improvement, such as clarifying instructions, vocabulary, and accessibility of materials, including in the RAICA Playground. Nearly 100 unique instances of revision opportunities were evaluated (accepted or rejected) by the curriculum development team. Finally, our study highlights the value of collaborating with educators to identify barriers and facilitators in classroom implementation. Teachers and learners serve as critical validators of AI curricula, and their insights help balance robustness and flexibility. While qualitative analysis of student projects was beyond this study’s scope, positive classroom experiences and open-ended student work suggest promising outcomes. DBIR proves essential in bridging the gap between curriculum design and real-world contexts, and contributing design principles that support pedagogically sound, equitable, and scalable AI literacy initiatives.

Acknowledgements

This project is funded by DP World and supported through the MIT RAISE initiative. We thank all our collaborators, including all teachers, students, and consulting AI experts.

References

American Institutes of Research. N.D. Considerations for Effective Implementation: 5 Elements of Fidelity. https://intensiveintervention.org/resource/five-elements-fidelity. Accessed: 2024-09-08.

Allison, C. 2023. Guidance Note on Using Implementation Research in Education, Building Evidence in Education (BE2). London: Foreign, Commonwealth & Development Office, prepared for Building Evidence in Education (BE2). https://inee.org/sites/default/files/resources/BE2%20IR%20guidance%20note%20Final%20version%202%20-%20March%202023_0.pdf Accessed: 2025-03-05.

Almatrafi, O.; Johri, A.; and Lee, H. 2024. A Systematic Review of AI Literacy Conceptualization, Constructs, and Implementation and Assessment Efforts (2019-2023). Computers and Education Open 6 (100173): 1-20. https://doi.org/10.1016/j.caeo.2024.100173

Brantmeier, E. J. (2005). Empowerment pedagogy: co-learning and teaching. Indiana University. http://www.indiana.edu/~leeehman/Brantmeier.pdf. Accessed: 2024-12-18.

Brennan, K., & Resnick, M. 2012. New Frameworks for Studying and Assessing the Development of Computational Thinking: Using Artifact-Based interviews to study the development of computational thinking in interactive media design. Paper presented at the 2012 American Educational Research Association meeting, Vancouver, BC, Canada, April 13-17.

Buxton, E.; Javadi, E.; Hagaman, M. 2024. Foundations of Autonomous Vehicles: A Curriculum Model for Developing Competencies in Artificial Intelligence and the Internet of Things for Grades 7–10. In Proceedings of the Association for the Advancement of Artificial Intelligence Conference. 38(21), 23276-23284. DOI:https://doi.org/10.1609/aaai.v38i21.30375.

Christensen, K. D. N., and West, R. E. 2018. The Development of Design-Based Research. In *Foundations of Learning and Instructional Design Technology: Historical Roots and Current* Trends, edited by Richard E. West. https://edtechbooks.org/lidtfoundations/development_of_design-based_research

Crompton, H. (2023). Evidence of the ISTE Standards for Educators leading to learning gains. *Journal of Digital Learning in Teacher Education 39*(4), 201–219. https://doi.org/10.1080/21532974.2023.2244089.

Fishman, B. J.; Penuel, W. R.; Allen, A.-R.; Cheng, B. H.; and Sabelli, N. 2013. Design-Based Implementation Research: An Emerging Model for Transforming the Relationship of Research and Practice. *National Society for the Study of Education* 112(2): 136–156.

Grover, S. 2021. Toward a Framework for Formative Assessment of Conceptual Learning in K-12 Computer Science Classrooms. In Proceedings of the 52nd Association for Computing Machinery Technical Symposium on Computer Science Education 31-37. doi.org/10.1145/3408877.3432460

Grover, S. 2024. Teaching AI to K-12 Learners: Lessons, Issues, and Guidance. In Proceedings of the 55th Association for Computing Machinery Technical Symposium on Computer Science Education V. 1 (Special Interest Group on Computer Science Education), March 20–23, 2024, Portland, OR, USA. https://doi.org/10.1145/3626252.3630937

Hemmeter, M. L.; Munson Doyle, P.; Collins, B. C.; and Jones Ault, M. 1996. Checklist for successful implementation of field-based research. *Teacher Education and Special Education* 19(4): 342-354. https://doi.org/10.1177/088840649601900406

Hill, D. R.; King, S. A.; Lemons, C. J.; and Partanen, J. N. 2012. Fidelity of Implementation and Instructional Alignment in Response to Intervention Research. *Learning Disabilities Research & Practice* 27(3): 116–124. https://doi.org/10.1111/j.1540-5826.2012.00357.x

Larmer, J., & Mergendoller, J. R. 2011. The Main Course, not Dessert. Buck Institute for Education.

Mishra, P.; Warr, M.; and Islam, R. 2023. TPACK in the Age of ChatGPT and Generative AI. *Journal of Digital Learning in Teacher Education* 39(4): 235-251. DOI: 10.1080/21532974.2023.2247480

Ng, D. T. K.; Leung, J. K. L.; Chu; S. K. W.; and Qiao, M. S. 2021. Conceptualizing AI literacy: An Exploratory Review. *Computers and Education: Artificial Intelligence* 2(100041): 1-11.

Olari, V., Tenório, K., Romeike, R. 2023. Introducing Artificial Intelligence Literacy in Schools: A Review of Competence Areas, Pedagogical Approaches, Contexts and Formats. In: Keane, T., Lewin, C., Brinda, T., Bottino, R. (eds) *Towards a Collaborative Society Through Creative Learning.* World Conference on Computers in Education 2022. IFIP Advances in Information and Communication Technology, vol 685. Springer, Cham. https://doi-org.libproxy.mit.edu/10.1007/978-3-031-43393-1_21

Papert, S. 1980. Mindstorms: Children, Computers, and Powerful Ideas. New York, NY: Basic Books.

Radday, E. & Mervis, M. 2024. AI, Ethics and Education: The Pioneering Path of Sidekick Academy. In Proceedings of the Association for the Advancement of Artificial Intelligence Conference. 38(21), 23294-23299. https://doi.org/10.1609/aaai.v38i21.30377

MIT. 2024. RAISE Playground: Interactive AI made Easy, Creative, and Fun. https://playground.raise.mit.edu/ Accessed: 2024-03-05.

Swain, M. S.; Finney, S. J.; and Gerstner, J. A. 2013. A Pratical Approach to Assessing Implementation Fidelity. *Assessment Update* 25(1): 5-17, 13. doi:10.1002/au

Su, J., and Yang, W. 2023. Artificial Intelligence (AI) literacy in early childhood education: an intervention study in Hong Kong. *Interactive Learning Environments* 32(9), 5494–5508. https://doi-org.libproxy.mit.edu/10.1080/10494820.2023.2217864

Tissenbaum, M.; Sheldon, J.; and Abelson, H. 2019. From Computational Thinking to Computational Action. *Communications of the Association for Computing Machinery* 62(3): 34–36.

Vahey, P.; Knudsen, J.; Rafanan, K.; and Lara-Meloy, T. 2013. Curricular activity systems supporting the use of dynamic representations to foster students' deep understanding of mathematics. In *Emerging technologies for the classroom: A learning sciences perspective*, edited by Chrystalla Mouza and Nancy Lavigne, 15-30. New York: Springer.

Wharton, S.; Gustafson-Quiett, M.C.; Bosch, C.A.; Davis, E.; Breazeal, C.; Abelson, H;, Klopfer, E. 2024. Responsible design: A design thinking process for students creating with AI. Poster presented at Play Make Learn Annual Conference. Madison, WI, July 18-29.

Yue, M.; Jong, M. S.; and Dai, Y. 2022. Pedagogical Design of K-12 Artificial Intelligence Education: A Systematic Review. *Sustainability* 14(15620): 1-29. https://doi.org/10.3390/su142315620