Supporting AI Literacy Teaching Through the Development of Assessments for Classroom Use

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

Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
j_masla@mit.edu, cabosch@mit.edu, prernar@mit.edu, lyd@mit.edu, wharton@mit.edu, marycate@mit.edu, samarh18@mit.edu, cmacatan@mit.edu, klopfer@mit.edu, cynthiab@media.mit.edu, hal@mit.edu

Abstract

Initial discussion of AI literacy assessment has focused on competency frameworks and learning standards rather than materials for classroom use. Responsible AI for Computational Action (RAICA), a constructionist AI curriculum for middle and high school students, includes assessment materials to support teachers with the evaluation of student AI literacy competencies in their classrooms. These materials include exit tickets used as formative assessments at the end of each lesson and both teacher and student-facing rubrics. After beta-testing a module of the curriculum with nine teachers and 282 students, we reviewed teacher usage data and feedback as well as student responses. The review process surfaced a number of improvements to the materials to better align them with classroom teaching practice. These included clarifying language and adding visual scaffolds. We present the assessment materials and iterative design process used to bridge the gap between the theoretical AI literacy competencies and their practical implementation in classrooms.

Introduction

The explosive growth of public attention on Artificial Intelligence (AI) in recent years has prompted the education community to consider how to best teach the technology in K-12 settings. This emerging field is referred to as AI literacy (Touretzky et al. 2019). Early AI literacy frameworks helped establish the field by synthesizing the complex domain of AI into a clear and organized set of competencies for students (Long and Magerko 2020; Ng et al. 2021; AI4K12 n.d.). These frameworks are an important contribution as they allow for the creation of educational activities and curricula with clear and relevant learning objectives.

Now that the AI literacy community has established *what* to teach, its attention turns to *how* to teach it. Assessment is a crucial component of quality classroom instruction in any discipline, as teachers use it to support student learning with feedback and to adjust their instruction (Young and Kim 2010; Black and Wiliam 2009). However, few AI literacy studies have focused on assessment in a classroom context. In a review of 32 AI teaching and learning papers, Yue, Jong, and Dai 2022 found that only 16% were conducted in regular class settings. Moreover, half were conducted in a single session, instead of over a series of lessons as would happen in a typical school schedule. AI assessment approaches often consist of multiple choice knowledge tests, self-reported attitude surveys, and project portfolio interviews (Ng et al. 2021), but "only a few studies (5 of 32) have developed a rubric or assessment sheet for evaluating students' understanding of what they have learned" (Yue, Jong, and Dai 2022). This indicates a gap in assessment materials that are practical for teacher use.

In order to support teachers with the implementation of AI literacy curricula, this study focuses on the development of assessments that inform teacher practice in a classroom context. We present a set of formative assessments and rubrics for an eight lesson module on the AI topic of image classification. These include exit tickets - formative assessments administered at the end of each lesson to give the teacher a quick gauge of student learning and support student metacognition - and both teacher-facing and student-facing rubrics. These assessments were developed to align with the AI4K12 learning guidelines and revised through review of the results of a beta test involving nine teachers and 282 students. In order to contribute to design-based research and curriculum development on how to teach AI in formal K-12 education settings, we share tools, processes and examples illustrating one approach to iteratively developing assessments that we invite others to apply and refine.

Background

AI Literacy Assessment

AI literacy assessments in the past have primarily used quantitative tools like pre- and post-knowledge tests to measure students' understanding of AI concepts (Lin and Van Brummelen 2021; Wan et al. 2020; Rodríguez-García et al. 2021). Other methods, such as project portfolio analysis and artifact-based interviews, provide insights into students' application of AI skills but lack standardized rubrics and are not scalable for typical classroom use due to time and resource constraints (Kaspersen et al. 2021; Kandlhofer and Steinbauer 2021; Ng et al. 2021; Zhang and Aslan 2021). Scholars have also pointed out that these efforts often do not focus on formative assessments or include teacher voices in the design process (Williams 2021; Ng et al. 2021). Another significant gap in AI literacy research is that most studies are

short-term (ranging from 3 hours to a week), researcher-led (Ravi et al. 2023; DiPaola, Payne, and Breazeal 2020; Burgsteiner, Kandlhofer, and Steinbauer 2016; Druga 2018; Lin et al. 2020), and conducted in extracurricular or online settings, limiting their relevance to regular classroom practice (Yue, Jong, and Dai 2022). While initiatives like AI4K12 have outlined foundational concepts, there is little emphasis on creating accessible, teacher-friendly tools and assessments that can be integrated into existing curricula. There is thus a need to iteratively develop formative assessment materials in collaboration with K-12 teachers, ensuring these materials are clearly aligned with competencies within AI literacy frameworks.

The RAICA Curriculum

The RAICA curriculum is designed to teach middle and high school students about AI through the creation of personally meaningful projects. It consists of five modules, each centered around an AI concept such as image classification or affective computing. Each module contains a teacher guide, slides, student handouts, and assessments. Rooted in the constructionist paradigm, the curriculum emphasizes learning through building, and seeks to position students to create with AI tools (Harel and Papert 1991). Students are empowered to engage in computational action by creating personally meaningful artifacts that may address real-world problems within their communities (Tissenbaum, Sheldon, and Abelson 2019). They do this through following the responsible design process, which builds on the design thinking process with an emphasis on stakeholders, impact and values (Wharton et al. 2024). Some examples of student projects are a poisonous vs. edible plant identifier, an image recognition app that offers advice on hair care and styling, and an interactive fish social robot that teaches about combatting water polution. Our study focuses on the development of assessments in the Picture This module of the RAICA curriculum. Through beta testing of the module in spring 2024 with nine teachers and 282 students, we gathered student response data, teacher use data, and teacher comments on the Picture This assessment materials. We used this data to revise the materials to better support teacher instruction.

Description of Resources

This resource includes tools – assessment materials, including exit tickets and two rubrics – and the processes used to iteratively design them with teacher feedback. These assessment materials provide an example of practical, classroom based tools that others building AI literacy curricula can adapt to their subject matter and learning objectives. Similarly, the iterative design process we outline here can be applied and refined by others seeking to incorporate teacher and student input. Before describing the assessment materials, we provide context about the learning module for which they were developed.

Picture This Learning Module

The Picture This module is centered around the AI concept of image classification. It begins with lessons to build knowledge of the technology, including the steps for creating an image classification model, the role of data, the potential for bias, and the basic function of a neural network. Students then apply this technical knowledge in the design of their own projects, in which they train image classification models and incorporate them into a program they build in the Scratch-like RAISE playground. In building projects with AI tools on personally relevant topics, students practice the skills of responsible design and computational action in addition to expanding their knowledge of AI.

Target Age Group Middle School students (ages 11-14)

Setup and Resources Needed The activities require at least one computer for every three students in a group, along with internet access. They will also need access to Google's Teachable Machine - a platform that enables the creation of image classification models (Google AI n.d.), and the RAISE Playground, which provides a Scratch-like environment with special extensions for integrating image classification models into their programs (MIT RAISE n.d.).

Expected Learning Outcomes The rubrics and formative assessments are designed to measure and support the attainment of the following learning outcomes:

Knowledge

- **Image Classification:** Explain how computers classify different categories of images from a dataset. *AI4K12 guideline 3-A-iii Nature of Learning (Training a model):* Train a classification model using machine learning, and then examine the accuracy of the model on new inputs.
- **Computer Perception:** Explain how computers use sensors to perceive the world. *AI4K12 guideline 1-A-ii Sensing (Digital Encoding):* Explain how images are represented digitally in a computer.
- **Neural Networks:** Explain the key components of neural networks and their role in image classification models. *AI4K12 guideline 3-B-i Neural Networks (Structure of a neural network):* Illustrate the structure of a neural network and describe how its parts form a set of functions that compute an output.
- **Data Bias:** Define data bias and articulate causes and remedies. *AI4K12 guideline 3-C-iii Datasets (Bias):* Explain how the choice of training data shapes the behavior of the classifier, and how bias can be introduced if the training set is not properly balanced.

Skills and Attitudes

- **Responsible Design:** Follow the steps of the responsible design process— Play, Brainstorm, Define, Pause & Plan, Prototype, and Try It Out.
- **Computational Action:** Increased self-efficacy and empowerment to design and build meaningful AI artifacts.

Assessment Materials

With this context, we now present the assessment materials and explain how they are designed to support instruction.

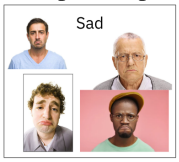
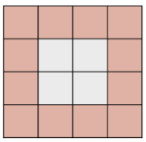
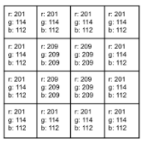
Lesson	Learning Objective(s)	AI4K12 Guideline	Questions
Part 1: Launch	- *Apply knowledge of image recognition software to demonstrate computer perception with Teachable Machine.* - *Develop familiarity with the tools used during this module (Teachable Machine & Scratch).*	**1-B-i:** Use a software tool such as a speech transcription or visual object recognition demo to demonstrate machine perception, and explain why this is perception rather than mere sensing.	1. Rate your comfort level using Scratch [1-5 Likert scale] 2. Rate your comfort level using Teachable Machine [1-5 Likert scale] 3. How would you describe Teachable Machine to someone unfamiliar with it?
Part 2: Image Classification	- *Define image classification and list the three steps of creating an image classification model.* - *Define data bias and articulate some causes and remedies of data bias in image classification.*	**1-C-i Domain Knowledge - Types:** Classify a given image and then describe the kinds of knowledge a computer would need in order to understand scenes of this type. **3-A-iii Nature of Learning (Training a model):** Train a classification model using machine learning, and then examine the accuracy of the model on new inputs.	1. Matching: Draw a line to order the steps used to create an image [Test the model, Train the model, Collect and label data; Step 1, Step 2, Step 3]    2. Given these classes, how would an image classification model classify the new image - Prediction for new image: Happy/Sad - Why do you think it would classify that way? 3. What are some potential sources of bias in this dataset?
Part 3: Computer Perception	- *Explain how images are processed by computers.* - *Contrast the strengths of humans and computers in processing images.*	**1-A-ii Sensing (Digital Encoding):** Explain how images are represented digitally in a computer.	1. Which of the following are valid RGB combinations? Circle all that apply. a. (178, 54, 54) b. (0%, 70%, 70%) c. (44, 625, -32)   Human perception Computer perception 2. Given these two examples, what is the difference between the way a human processes images and a computer processes images? 3. What do you as a human see in this image? What would a computer ”see”? Would a human or computer be better at knowing that this is a picture of a cat? Why? 

Table 1: Exit Tickets Parts 1-3

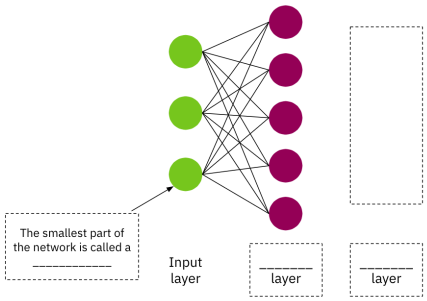
Lesson	Learning Objective(s)	AI4K12 Guideline	Questions
Part 4: Neural Networks	*Explain what happens in a neural network when an image classification model is being trained.*	**3-B-i Neural Networks (Structure of a Neural Network):** illustrate the structure of a neural network and describe how its parts form a set of functions that compute an output.	1. Complete and label the diagram of a neural network  2. What does a neuron do in a neural network? a. Predicts the weather b. Stores memories c. Receives and processes input 3. How would you explain neural networks to a friend?
Part 5: Ideate and Define	*- Identify a focus for a project.* *- Plan methods for including stakeholder input.*	**3-A-iii Nature of Learning (Training a Model):** Train a classification model using machine learning, and then examine the accuracy of the model on new inputs.	1. What ideas did you brainstorm/ideate for your project? 2. Who are your stakeholders? How will your project make an impact on your stakeholders? How do you plan to get feedback from your stakeholders? 3. Rate your agreement with the following statement: My project is meaningful to me. [Likert scale 1-5]
Part 6: Prototype	*Create a working prototype of the project that integrates Teachable Machine and Scratch to address stakeholder needs.*	**3-C-i Datasets (Feature Sets):** Create a labeled dataset with explicit features of several types and use a machine learning tool to train a classifier on this data.	1. Check the project rubric. What is an area your project does well? 2. Check the project rubric. What is an area you want to improve for your project? 3. What kinds of problems or challenges do you have when coding with Scratch?
Part 7: Try It Out	*Test and iterate project to align with stakeholder needs.*	**3-A-iii Nature of Learning (Training a Model):** Train a classification model using machine learning, and then examine the accuracy of the model on new inputs.	1. What was one piece of feedback you got about your project? 2. What did you change based on that feedback? 3. Rate your agreement with the following statement: I consider myself part of the AI-design community. [Likert scale 1-5]
Part 8: Showcase	*Demonstrate learning through presenting work to community members.*	**3-A-ii Nature of Learning (Finding Patterns in Data):** Model how supervised learning identifies patterns in labeled data. **3-C-II Datasets (Large Datasets):** Illustrate how training a classifier for a broad concept such as ”dog” requires a large amount of data to capture the diversity of the domain.	What was the most important thing you learned about AI during this unit?

Table 2: Exit Tickets Continued Parts 4-8

picture this | teacher-facing rubric

	Below Standard	Approaching Standard	At Standard
Purpose	Project's purpose is unclear or not connected to stakeholders.	Project purpose is defined, but may not fully address stakeholder needs.	Project has a clearly defined goal that is relevant to stakeholders.
Stakeholder Engagement	No consideration of harms & benefits to stakeholders. No evidence of stakeholder input.	Some consideration of harms & benefits to stakeholders. Partial evidence of stakeholder input.	Clear consideration of harms & benefits to stakeholders. Extensive evidence of stakeholder input.
Accuracy	Image classification model is not present or has major flaws that significantly impact the function of project.	Image classification model may be accurate for some users, but this group is so narrow that project does not achieve purpose.	Image classification model is accurate for most users & functions smoothly enough to contribute to project purpose.
Code	Code does not integrate image classifier into project.	Code integrates image classifier into project but lacks clarity or does not run smoothly.	Code successfully integrates image classifier into project, runs smoothly & contributes to project purpose.
Technical Explanation	No explanation of computer perception & NN included in presentation.	Limited explanation of computer perception & NN included in presentation.	Clear explanation of computer perception & NN, how they impact project function.

Table 3: Teacher-Facing Rubric

Formative Assessments These assessments consist of exit tickets designed to give teachers a quick gauge of student learning in each lesson's content and to support student metacognition. Teachers can then use this information to offer additional scaffolds and adjust their instruction. For example, teachers may pull a small group of students in the following lesson to review a misunderstood concept. Each exit ticket consists of no more than three questions, with a mix of open-response, multiple choice, and Likert scale. See **Tables 1 and 2** for the exit tickets and their associated learning objectives, mapped to the AI4K12 guidelines they cover.

Rubrics The module contains teacher-facing and student-facing rubrics. The rubrics were designed with guidance of PBLWorks "Rubric for Rubrics" (PBLWorks 2019), whose recommendations include starting with clear criteria, creating meaningful distinction between levels of achievement, and using student-friendly language. The criteria, developed in alignment with the learning objectives of the module, are: Purpose, Stakeholder Engagement, Accuracy, Code, and Technical Explanations. The rubric contains three levels of achievement (At Standard, Approaching Standard, and Below Standard) described in student-friendly language. See the teacher-facing rubric in **Table 3**.

The student-facing rubric is intended to serve as a learning scaffold and formative assessment. It is formatted as a single-point rubric (Gonzalez 2014), including only the description of the "At Standard" category in a central column. On either side of the description, there is space for students to write evidence that their project exceeds the description or to identify areas of improvement. This encourages students to see their project as dynamic and open to iteration, as opposed to the typical static judgment of a "finished product" that comes with analytical rubric scores (Fluckiger 2010).

Iterative Design and Revision Process

This section outlines how our assessment materials were revised using teacher feedback and student data to enhance classroom instruction, providing a process for future scholars to use and adapt when developing their own AI learning curricula.

Participants In the Spring of 2024, the Picture This module was beta tested by 9 teachers across 4 continents, with a total of 282 students. The study sample included computer science teachers and students from private schools, public schools, and an education non-profit in the United States, Chile, Malawi and U.A.E. The study was approved by MIT's internal review board (IRB). During the spring beta test, we piloted the method of teacher interview with one teacher. We interviewed the teacher before and after implementing the curriculum, including questions about the teacher's pedagogy, AI content knowledge, and assessment practices. This data was used to complement findings from the other data sources, but not as primary evidence.

Formative Assessments

Data Collection Participating teachers provided usage data and feedback by filling out a fidelity tracker as they progressed through the module. In addition to feedback on lesson activities and teacher notes, for each assessment, they were asked "How did you use this activity?" with the choices: "A. Exactly as prescribed B. Modified activity C. Modified timing D. Modified student grouping E. Replaced completely F. Skipped." They were also prompted to "Please explain any modification or replacements." Teachers provided student data by uploading student responses to exit tickets into a folder accessible by the research team. Different implementation rates and data collection practices led to a wide range in the number of student responses for each exit ticket. The highest number of student responses was 121 and the lowest was 19. We did not receive any data on the eighth and final exit ticket, as teachers incorporated it into the student presentation activities, which removed the opportunity for data collection.

Data Analysis and Revision The research team analyzed each exit ticket item using a summary of student responses to guide their revision process. For multiple choice items, we used the mean of correct responses. For Likert scale items, we summarized with mean, median and mode. For open response items, we reviewed, highlighted key terms, and identified emerging trends across the sample of responses.

The next step was to analyze the data through the lens of teacher instruction and student learning. The research team reviewed the summary data with the following questions in

mind: *(1) Does this item work to give teachers a quick gauge of student learning? Can it help inform teacher instruction? (2) What, if anything, can we infer about student learning based on this data? (3) What should we be learning about our work from this item?* After engaging in a generative discussion of these questions, the team synthesized their insights into concrete ideas for revising the item or elements of the curriculum related to the item.

After completing the exit ticket analysis document, the team compared generated ideas for revision with teacher usage, comments, and feedback. Teacher comments about their modifications to exit tickets were crucial in determining final revisions.

After generating revisions based on student responses, teacher usage data, and comments, the team organized these insights through coding the formative assessment items. **Figure 1** provides an overview of this process.

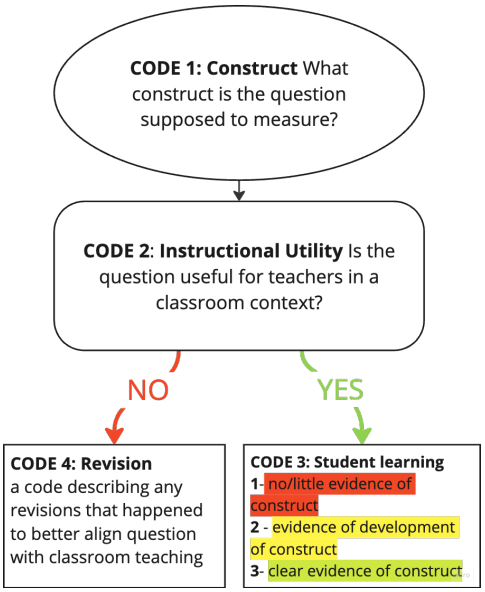


Figure 1: Process for analyzing exit tickets

Code 1: Construct - *What construct is the question supposed to measure?* This was determined by reviewing the learning objectives of the lesson and the text of the question. Codes for this category were: AI literacy– data bias, AI literacy– image processing, AI literacy– neural network, Responsible Design, Responsible Design– stakeholders, and Computational Action.

Code 2: Instructional Utility - *Is the question useful for teachers in a classroom context?* This was determined by analyzing teacher usage and student response data. If teachers did not use the item or heavily modified it, the item was likely not sufficiently useful to their instruction. If students seemed to misunderstand the question (or the question did not generate specific enough answers for the construct to be assessed), the question was not useful generating data about their learning of the construct. The codes for this category were "yes" and "no."

Code 3: Student Learning - *Does the exit ticket produce evidence of student learning?* This code was applied to the portion of exit ticket items determined to have instructional utility. To compare student learning across question types, we created a simple rubric: 1 = little or no evidence of construct; 2 = evidence of emerging development of construct; 3 = clear evidence of construct. We reviewed student responses and coded them according to this rubric.

Code 4: Revision - *What revisions were made to the item?* This code was applied to the portion of exit ticket items that were determined to not have instructional utility. These were developed inductively based on revisions to the exit tickets made by the research team after considering student responses and teacher feedback. The codes were: *made question more specific, clarified vocabulary, clarified question, modified question format, removed or replaced question,* and *added visual.*

Rubrics

Data Collection Data on rubrics included teacher feedback and a selection of eight student projects. The researchers asked participating teachers to submit example student projects to a student showcase, which yielded a total of eight projects. The data teachers submitted included the Scratch-like code, a copy of the slideshow students created to present their work, and the graphic organizers students used to facilitate the creation of their projects.

Revision Process The research team used the analytical rubric to evaluate submitted student projects. They assigned a score for each category and wrote a description of why they assigned that score. Using the rubric tool to review actual student work generated insights and ideas for revision.

Findings

We found that 13 of the 20 exit ticket items had instructional utility. For these 13 items, we then coded student response data using the rubric for evidence of student learning. Nine of these items were determined to have clear evidence of learning the construct and four as having evidence of developing understanding of the construct. None of the exit tickets determined to have instructional utility were coded as having no or little evidence of the construct. This indicates that these 13 exit ticket items generated insights on student learning.

In addition to the 7 items determined not to have instructional utility, we surfaced 6 revisions for items determined to have instructional utility. In these cases, the item was deemed useful for measuring student learning of the construct but still benefited from changes. **Figure 2** shows the revision codes and their frequency.

Teacher feedback on the rubrics was generally positive. One said "I like the provided rubric and the kids understood it," indicating the language and layout of the rubric was accessible to their students. Another appreciated having the option of using either the student-facing single-point rubric or the teacher-facing analytical rubric. In an interview, a teacher described that they found the rubric "really helpful, the table, the different categories. It was readable, accessible for students… It was good when they assessed their peers

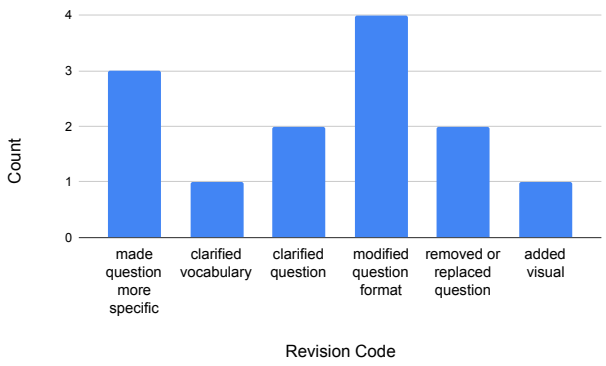


Figure 2: Revision Codes

and filled in feedback. We did another activity with informal feedback." They not only appreciated the clarity of the rubric, but also its ability to be used for peer feedback.

In addition to peer feedback, the student-facing rubric was used for self-reflection during project building. At the end of a prototyping lesson 6, the exit ticket asks students to review the rubric to identify an area of strength and improvement for their project. Of the students who used the rubric during this activity (n=98), 88% gave a response involving one of the five rubric categories. This indicates the rubric supported reflection on their project building skills.

The internal review of projects using the rubrics indicated they are useful for evaluation. The research team was able to find evidence of student learning for each category and choose a score based on the described criteria. A revision to the rubric also surfaced from this process, which was to make the "Stakeholder Engagement" category criteria more specific by not only having a plan to gather feedback, but to also incorporate it into project design.

Discussion

Design Insight 1: Clarity of Language

Clarity of language emerged as a design insight when reviewing both formative assessment and rubric data. In formative assessment revision, the codes "made question more specific," "clarified vocabulary," and "clarified question" made up 46% of the total. Teachers reported appreciation for the clear, accessible language in the rubrics.

Unlike more well-known technologies, the new and rapidly emerging field of AI is filled with domain-specific terms that may be unfamiliar to the general public. These terms are the most difficult for new readers and require intentional instruction to lower barriers to understanding (Flanigan and Greenwood 2007). The field of AI education may be particularly jargon-filled due to its roots in a highly technical subject previously only taught at the university level. Creators of AI education materials should be especially aware of the use of domain-specific words in their learning materials, grounding students with specific examples and straight-forward language.

Design Insight 2: Visual Scaffolds

Another design insight was to provide visual scaffolds in the formative assessment materials. For example, the exit ticket in lesson four originally asked students to draw a neural network and provided only a blank box. Only three of the nine teachers used this assessment as described. The teachers who modified it all provided a diagram for their students. Taking this into account, we revised the item to include a partially complete diagram with labels to key features of neural networks, and asked students to complete and label the diagram.

The use of visual scaffolds, images and graphic organizers, is a proven instructional method in other educational contexts (Gallavan and Kottler 2007; Kim et al. 2004). They reduce cognitive load by providing structure for students to map their understanding (Park 2017). Our work with teachers indicates a preference for including visual scaffolds in formative assessments in AI education materials. This strategy should be embraced to support students in representing their emerging knowledge of AI.

Limitations

Due to the sample size and specific classroom contexts, these findings are not generalizable. When teachers submitted a selection of student projects in data collection, they may have only chosen what they considered the "best" projects, implying projects further from standard were not in the sample, influencing findings on rubric utility. There was a wide variety of responses collected for exit tickets, with a low of 19 and a high of 121. The conclusions drawn from the exit tickets with fewer student responses may be less representative. Future work should explore the long-term impacts of these assessment materials on student understanding and teacher pedagogy.

Conclusion

In this paper, we present a practical approach to assessing AI literacy in middle school classrooms. This includes examples of formative assessments and rubrics for a constructionist module on the AI topic of image classification. We also outline the process used to generate these assessments through considering teacher feedback and student response data. We surface insights on practical classroom AI literacy assessments, including clarity of language and use of visual scaffolds. Both the assessment tools and the process used to create them can serve as examples for those seeking to implement AI education in a classroom context.

Acknowledgments

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

References

AI4K12. n.d. AI4K12 Website. Last accessed September 16, 2024.

Black, P.; and Wiliam, D. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of personnel evaluation in education)*, 21: 5–31.

Burgsteiner, H.; Kandlhofer, M.; and Steinbauer, G. 2016. Irobot: Teaching the basics of artificial intelligence in high schools. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

DiPaola, D.; Payne, B. H.; and Breazeal, C. 2020. Decoding design agendas: an ethical design activity for middle school students. In *Proceedings of the interaction design and children conference*, 1–10.

Druga, S. 2018. *Growing up with AI: Cognimates: from coding to teaching machines*. Ph.D. thesis, Massachusetts Institute of Technology.

Flanigan, K.; and Greenwood, S. C. 2007. Effective content vocabulary instruction in the middle: Matching students, purposes, words, and strategies. *Journal of Adolescent & Adult Literacy*, 51(3): 226–238.

Fluckiger, J. 2010. Single point rubric: A tool for responsible student self-assessment. *The Delta Kappa Gamma Bulletin*, 76(4): 18.

Gallavan, N. P.; and Kottler, E. 2007. Eight types of graphic organizers for empowering social studies students and teachers. *The Social Studies*, 98(3): 117–128.

Gonzalez, J. 2014. Know Your Terms: Holistic, Analytic, and Single-Point Rubrics. Accessed December, 2024.

Google AI. n.d. Teachable Machine. Last accessed September 16, 2024.

Harel, I. E.; and Papert, S. E. 1991. *Constructionism.* Ablex Publishing.

Kandlhofer, M.; and Steinbauer, G. 2021. AI K–12 education service. *KI-Künstliche Intelligenz*, 35: 125–126.

Kaspersen, M. H.; Bilstrup, K.-E. K.; Van Mechelen, M.; Hjorth, A.; Bouvin, N. O.; and Petersen, M. G. 2021. VotestratesML: A high school learning tool for exploring machine learning and its societal implications. In *FabLearn Europe/MakeEd 2021-An international conference on computing, design and making in education*, 1–10.

Kim, A.-H.; Vaughn, S.; Wanzek, J.; and Wei, S. 2004. Graphic organizers and their effects on the reading comprehension of students with LD: A synthesis of research. *Journal of learning disabilities*, 37(2): 105–118.

Lin, P.; and Van Brummelen, J. 2021. Engaging teachers to co-design integrated AI curriculum for K-12 classrooms. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–12.

Lin, P.; Van Brummelen, J.; Lukin, G.; Williams, R.; and Breazeal, C. 2020. Zhorai: Designing a conversational agent for children to explore machine learning concepts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13381–13388.

Long, D.; and Magerko, B. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–16.

MIT RAISE. n.d. RAISE Playground. Last accessed September 16, 2024.

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.

Park, S. 2017. An exploratory study on the meaning of visual scaffolding in teaching and learning contexts. *Educational Technology International*, 18(2): 215–247.

PBLWorks. 2019. Rubric for Rubrics. Accessed September 16, 2024.

Ravi, P.; Broski, A.; Stump, G.; Abelson, H.; Klopfer, E.; and Breazeal, C. 2023. Understanding Teacher Perspectives and Experiences after Deployment of AI Literacy Curriculum in Middle-school Classrooms. *ICERI2023 Proceedings*, 6875–6884.

Rodríguez-García, J. D.; Moreno-León, J.; Román-González, M.; and Robles, G. 2021. Evaluation of an online intervention to teach artificial intelligence with learningml to 10-16-year-old students. In *Proceedings of the 52nd ACM technical symposium on computer science education*, 177–183.

Tissenbaum, M.; Sheldon, J.; and Abelson, H. 2019. From computational thinking to computational action. *Communications of the ACM*, 62(3): 34–36.

Touretzky, D.; Gardner-McCune, C.; Martin, F.; and Seehorn, D. 2019. Envisioning AI for K-12: What should every child know about AI? In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 9795–9799.

Wan, X.; Zhou, X.; Ye, Z.; Mortensen, C. K.; and Bai, Z. 2020. SmileyCluster: supporting accessible machine learning in K-12 scientific discovery. In *Proceedings of the interaction design and children conference*, 23–35.

Wharton, S.; Gustafson-Quiett, M.; Bosch, C.; Davis, E.; Breazeal, C.; Abelson, H.; and Klopfer, E. 2024. Responsible design: A design thinking process for students creating with AI. In *Play Make Learn Annual Conference*. Madison, WI, United States. Poster session.

Williams, R. 2021. A Review of Assessments in K-12 AI Literacy Curricula. Accessed December, 2024.

Young, V. M.; and Kim, D. H. 2010. Using assessments for instructional improvement: A literature review. *Education policy analysis archives*, 18: 19–19.

Yue, M.; Jong, M. S.-Y.; and Dai, Y. 2022. Pedagogical design of K-12 artificial intelligence education: A systematic review. *Sustainability*, 14(23): 15620.

Zhang, K.; and Aslan, A. B. 2021. AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*, 2: 100025.