International Society of the Learning Sciences

Design Considerations for Evaluating Middle School AI Knowledge

Xueru Yu, Georgia Institute of Technology, xyu425@gatech.edu
William Gelder, Georgia Institute of Technology, wgelder3@gatech.edu
David Touretzky, Carnegie Mellon University, dst@cs.cmu.edu
Christina Gardner-McCune, University of Florida, gmccune@ufl.edu
Judith Uchidiuno, Georgia Institute of Technology, jiou3@gatech.edu

Abstract: K-12 students need to understand technical and ethical knowledge about Artificial Intelligence to utilize ubiquitous AI-powered technologies responsibly. Recent studies have explored ways to teach AI to K-12 students effectively, but little is known about ways to assess their learning. In this study, we describe how five computer science teachers from urban and rural Georgia school districts designed, adapted, and implemented assessments in their classrooms while teaching a middle school elective course aligned with the Five Big Ideas of AI. Analyzing artifacts from 201 students, we explore the efficacy of different assessment instruments co-designed with teachers, measuring students' understanding of sensors and the societal impacts of autonomous vehicles. We suggest design considerations for AI knowledge assessments to meet the needs and challenges in diverse classroom contexts.

Introduction

Artificial Intelligence (AI) is increasingly prevalent in various aspects of youth's everyday lives. They regularly interact with AI-powered technologies such as voice assistants and smart home devices, and their parents may own semi-autonomous vehicles. Teaching AI to K-12 students empowers them to use AI knowledgeably and create socially responsible AI-powered tools in the future.

Several AI researchers and educational stakeholders are developing K-12 AI curricula, covering a set of technical knowledge (e.g., what AI is), and ethical knowledge (e.g., how automated reasoning algorithms may fail to treat people equitably) (Kong et al., 2024; Long & Magerko, 2020; Williams et al., 2023). However, development of AI assessments has not kept pace with growing AI curricula. Formative and summative assessments measure students' learning and inform the alignment between curriculum, pedagogies, and desired learning outcomes. While existing literature reviews have identified commonly used AI assessments, such as performance projects (Marques et al., 2020; Ng et al., 2021), there is limited understanding about how to design these assessments to reflect mastery of key AI concepts accurately. Additionally, assessments should give all students equal opportunities to demonstrate what they know. Ko et al. (2024) have proposed principles for equitable assessments in CS education, yet what they look like in AI education remains under-explored.

In response, we co-designed assessments used in a middle school AI elective with CS teachers from five school districts in Georgia, USA. Co-design is a process in which stakeholders work collaboratively in defined roles and provide equal input to realize innovations (Roschelle et al., 2006). Since we involved teachers as equal design partners, the resulting assessments reflected teacher perspectives of what effective assessments look like in their classrooms, regardless of their actual efficacy for uncovering students' AI knowledge. We explored the following research questions: (1) *What assessments do teachers deem effective for assessing students' AI knowledge, such as their understanding of sensors, and the societal impacts of autonomous vehicles (AV) and autonomous robots (AR)*? (2) *To what extent did the co-designed assessments demonstrate the targeted learning objectives?* We focused on the two knowledge areas because understanding sensors is an example of technical AI knowledge and is fundamental to understanding how AI devices perceive and react to the world. Understanding societal impacts is an example of ethical AI knowledge that directly addresses responsible AI use.

Theoretical framing

Design-based research (DBR) is a methodology aimed at creating educational interventions and improving theories of teaching and learning within real-world contexts. Reeves (2006) suggested four phases of design-based research: (a) analysis of practical problems by researchers and practitioners, (b) development of solutions with a theoretical framework, (c) evaluation and testing of solutions in practice, and (d) documentation and reflection to produce "design principles." Researchers iterate through these phases to continuously refine their theories and practices. Using the DBR approach , we co-designed assessments used in a middle school AI elective with CS teachers and implemented them at schools. This study represented phases (c) and (d), where we evaluated the efficacy of assessments to improve our assessments in subsequent iterations and guide research on AI assessments.

Methods

The AI4GA project is a three-year initiative to develop a nine-week AI elective course (1) for Georgia middle school students. Aligned with the Five Big Ideas of AI (Touretzky et al., 2019), the course leaves room for teachers to adapt materials. This study focused on students' learning from Unit 1: Autonomous Robots & Vehicles. This unit primarily used unplugged activities to introduce concepts, including sensors, breadth-first search, and societal impacts of AI. It culminated in a design-based project called My Dream Bot, where students imagined an AR that solves a societal problem they care about. At the end of each module, we provided 'exit ticket' prompts that divided the project into manageable parts and guided students to add more details about how their robot works based on the main concepts presented in the module.

During 2022-2023, we recruited 5 middle school CS teachers as co-designers (Gardner-McCune et al., 2023), representing 5 school districts across 2 ethnically and geographically diverse populations in Georgia: rural and urban areas (Table 1). Teachers were compensated 1,000 US dollars for participating in the weekly meetings and collecting student assessment data. Each teacher taught two sections of students, lasting between 45 and 70 minutes, and either taught every day or every other day in a block schedule. This IRB-approved study was conducted with the consent/assent of parents/guardians, students, teachers, and school districts.

We collected student artifacts and interviewed teachers at the end of the school year. Each teacher interview was video recorded, lasted for about one hour, and consisted of questions about their adaptations of curriculum materials and their experiences with co-design. We conducted artifact analysis to evaluate the correctness of students' responses. For open-ended questions, we coded students' responses using emergent categories, which were iteratively refined (Clarke & Braun, 2017). We also transcribed teacher interviews and examined teachers' rationale for modifying each assessment.

Table 1
Student Demographic Information

Teacher ID	District Type	Student Demographic	# Students	% Students Eligible for Free/Reduced Lunch	% Students Who Have Taken Prior Computing Courses
A	Rural	Mostly White and Black	62	93%	100%
B	Urban	Mostly Black and White. Students enrolled and not enrolled in the International Baccalaureate program took our course together.	50	80%	85%
C	Rural	Mostly Black and White	29	95%	70%
D	Urban	Mostly Black	20	100%	67%
E	Urban	Mostly Black, some Hispanic	40	91%	49%
Total			201		

Results

Learning about sensors in autonomous vehicles and robots

In this module, students learned to describe the locations and functions of sensors (camera, LIDAR, RADAR, GPS) commonly found in/on AVs and ARs. Teachers provided slides with definitions, discussion prompts, and videos of a Nuro R2 vehicle navigating city streets. Students completed graphic organizers and illustrated diagrams to label sensors on the Nuro R2 vehicle.

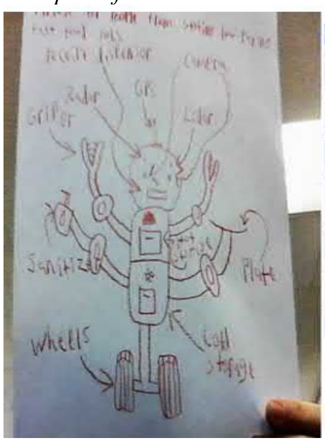
To assess these learning goals, teachers co-designed with the research team (1) an end-of-unit activity that had students create a one-page advertisement for an AV and explain sensors and their locations (used by teacher A); and (2) an exit ticket as part of the My Dream Bot project regarding the things their AR will sense, sensors used, and locations of all sensors (used by teacher A, B, C).

The assessments were adaptable. For the My Dream Bot project, teacher C felt that her students needed more support, so she developed sample responses with a support robot for people with physical disabilities and described its sensors: "*First semester, we didn't have that slide in there…I got some of everything that didn't make sense. But on the second semester, we added that slide in there with the robot for it. It really led to them some structural thinking. They were able to answer …like where would the sensors go*." However, not all teacher adaptations were effective. Teacher D's students struggled with reading and writing in other subjects. To accommodate them, teacher D removed AI knowledge questions, including sensor ones, and had students create a slideshow to show what their robots looked like and explain their functions.

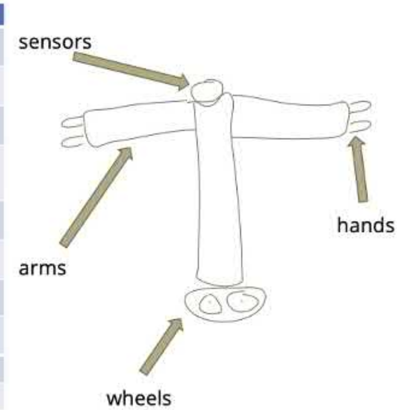
In both assessments, designing AI agents overshadowed the application of sensor knowledge, undermining their efficacy in capturing students' understanding. In the one-page advertisement, 18% named sensors but did not identify any sensor position. They developed elaborate market plans for their AV, describing them as unusual, durable, highly automated, and always safe, often choosing images of luxury cars to represent the product. In the My Dream Bot project, 37% used cameras without describing functions, 37% for LIDAR, 35% for RADAR, and 29% for GPS. They added many functional features, such as storage compartments and extendable arms for reaching high places, or incorporated enhanced capabilities, such as increased strength and speed.

For students who engaged with sensor questions, the assessments still provided valuable insights into their learning. Most students were able to explain sensor functions. In the My Dream Bot project, 58% described camera functions correctly, 53% described LIDAR functions correctly, 58% described RADAR functions correctly, and 57% described GPS functions correctly. However, although their ARs were designed with distinct functions and operated in specific environments, students did not specify what things the camera, LIDAR and RADAR sense (e.g., traffic signs, pedestrians, buildings). Most students were also able to explain sensor placement. In the one-page advertisement and the My Dream Bot project respectively, 50% and 66% described at least one sensor location correctly. Common knowledge gaps were placing object detection sensors such as RADAR inside AVs and ARs or placing all sensors on ambiguously stated positions. Figure 1 (b) used cameras, motion, and sound sensors and labeled all sensors on the robot's "head".

Figure 1
Examples of Student Dream Bot (a) Sensors Labeled Correctly (b) Sensors Labeled Ambiguously



Feature	Description	Location
GPS	Shows the position of Burger Bot.	Top of Robot
Radar	Shows LARGE objects at CLOSE range.	Top Left Eye
Lidar	Uses light to create a 3D environment.	Top Right Eye
Cameras	Shows a colored Image of the environment and reads table signs.	Bottom set of Eyes
Storage compartments	Keeps food at its correct temperature.	Torso
Grippers	Allows Burger bot to manipulate objects.	Top arms
Sanitizer	Allows Burger bot to clean surfaces.	Bottom Left Arm
Plate	Allows burger bot to carry many plates at a time.	Bottom Right Arm
Wheels	Allows Burger Bot to move	Bottom of Robot
Receipt Dispenser	Prints anything that's needed	Mouth



Learning about societal impacts of autonomous vehicles and robots

In this module, students learned to describe the positive and negative impacts of AVs and ARs and consider multiple stakeholders in evaluating their impacts. Teachers created slides with videos reporting Tesla drivers asleep at the wheel on a bridge during rush hour, and explanations of different automation levels. Students discussed ethical issues related to the liability of drivers, manufacturers, and lawmakers, and debated about appropriate policies to regulate the use of AVs.

To assess these learning goals, teachers co-designed with the research team: (1) Google Slide discussion workbooks with questions such as "should autopilot mode be mode" (used by teacher A, B, C); (2) an exit ticket as part of the My Dream Bot project regarding concerns people might have about their AR, aspects of ARs that have positive impacts on society, workplaces that might be interested in using their AR (used by teacher A, B, C).

The assessments were also adaptable. For the discussion workbook, teacher C anticipated that students have a short attention span and used an interactive learning tool PearDeck that turned questions into polls and discussion boards to increase engagement. For the My Dream Bot project, teacher B's students had learned the problem-solving and design thinking process in computing prior to our AI course. Teacher B connected to this prior instruction and started the project by having students identify the problems they wanted to solve and the stakeholders they targeted, rather than leaving them until the end.

The discussion workbook effectively assessed students' understanding of AV's societal impacts. When debating whether or not autopilot mode in AVs should be banned, 72% of students argued against a ban, citing benefits for different groups like the elderly, people with disabilities, novice drivers, and drivers who need to multitask. One student shared that her diabetic father needed to take medication during emergencies: "*My dad is diabetic, and you never know when your blood sugar will drop and you need to eat something*." A few students also suggested solutions like alert sounds when drivers' hands leave the wheel or higher fines for misuse. 25% of students argued that autopilot should be banned due to safety risks. 3% of students did not express a clear stance and presented arguments for both sides.

In contrast, the My Dream Bot project did not elicit a thoughtful and grounded evaluation of AR's societal impacts. Many students envisioned AR as their personal assistant to handle chores and prepare meals, or support industries like healthcare and customer service. When asked about their AR's broader societal impacts, only 24% of students responded. Many mentioned that it might make people lazy; some were influenced by common fictional plots and concerned that their AR would become evil or take over the world, despite its intended use for everyday tasks.

Discussion

This work analyzed assessments related to sensors and societal impacts about AVs and ARs used in a middle school AI literacy curriculum. We presented how teachers used and adapted assessments for their students. The one-page advertisement and My Dream Bot project could be improved to ensure students cover and contextualize the targeted AI knowledge within their AI agents while encouraging creativity.

Leaving room for teacher input makes AI assessments more equitable. Teachers will encounter students with varying CS interests, prior knowledge, and literacy competencies. Some teachers like teachers A and B may feel comfortable with the curriculum pace and have opportunities to implement additional assessments or make existing assessments more cognitively demanding. Some teachers like teacher C may only have time for covering core assessments embedded in modules. This finding is aligned with prior studies showing that teachers found customizable AI curriculum that allowed selective integration of modules to be helpful (Lee & Perret, 2022; Walsh et al., 2023). Researchers can provide teachers with both mandatory assessments and optional ones that allow in-depth explorations of curriculum topics to cater their needs and pace.

Teachers sometimes oversimplify assessments to accommodate students with a perceived struggle. Teachers like teacher D may use strategies that eliminate the written aspects of assessments related to AI knowledge and not replace them with alternative tasks. Such strategies compromise the depth of assessments. Prior studies have demonstrated examples of successful teacher adaptations that better meet the needs of students (Kim et al., 2021; Lin & Van Brummelen, 2021). Our work built upon these studies by showing that not all adaptations were effective. Researchers can develop and implement professional development workshops to address AI teachers' assessment literacy.

As AI education continues to evolve, it is crucial to develop robust assessments that can gauge diverse students' understanding about AI. This paper highlights the need for improving AI teachers' assessment literacy and discusses the affordances and challenges with our assessment practices. We also propose recommendations for AI projects to strike a balance between encouraging creative expressions and evaluating AI knowledge. We encourage both researchers and educators to attend to these aspects when integrating assessment literacy into teacher PD and in the classroom in order to improve the quality of K-12 AI education and assessment.

Endnotes

(1) Full curriculum link: https://sites.google.com/view/ai4gayear2resources/curriculum/Overview

References

Clarke, V., & Braun, V. (2017). Thematic analysis. *The Journal of Positive Psychology*, *12*(3), 297–298. https://doi.org/10.1080/17439760.2016.1262613

Gardner-McCune, C., Touretzky, D., Cox, B., Uchidiuno, J., Jimenez, Y., Bentley, B., Hanna, W., & Jones, A. (2023). Co-Designing an AI Curriculum with University Researchers and Middle School Teachers. *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2*, 1306. https://doi.org/10.1145/3545947.3576253

Kafai, Y. (1995). *Making Game Artifacts To Facilitate Rich and Meaningful Learning*. https://eric.ed.gov/?id=ED388682

Kafai, Y. (2012). *Minds in Play* (1 ed.). Routledge. https://doi.org/10.4324/9780203052914

Kim, Y. J. (Yj), Murai, Y., & Chang, S. (2021). Implementation of embedded assessment in maker classrooms:

Challenges and opportunities. *Information and Learning Sciences*, *122*(3/4), 292–314. https://doi.org/10.1108/ILS-08-2020-0188

Ko, A. J., Beitlers, A., Wortzman, B., Davidson, M., Alannah, O., Kirdani-Ryan, M., Druga, S., & Everson, J. (2024). *Critically Conscious Computing: Methods for Secondary Education*. criticallyconsciouscomputing.org

Kong, S.-C., Cheung, M.-Y. W., & Tsang, O. (2024). Developing an artificial intelligence literacy framework: Evaluation of a literacy course for senior secondary students using a project-based learning approach. *Computers and Education: Artificial Intelligence*, *6*, 100214. https://doi.org/10.1016/j.caeai.2024.100214

Lee, I., & Perret, B. (2022). Preparing High School Teachers to Integrate AI Methods into STEM Classrooms. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12783–12791. https://doi.org/10.1609/aaai.v36i11.21557

Lin, P., & Van Brummelen, J. (2021). Engaging Teachers to Co-Design Integrated AI Curriculum for K-12 Classrooms. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3411764.3445377

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727

Marques, L. S., Gresse Von Wangenheim, C., & Hauck, J. C. R. (2020). Teaching Machine Learning in School: A Systematic Mapping of the State of the Art. *Informatics in Education*, 283–321. https://doi.org/10.15388/infedu.2020.14

Reeves, T. (2006). Design research from a technology perspective. In *Educational Design Research*. Routledge.

Roschelle, J., Penuel, W. R., Shechtman, N., & Penuel, W. (2006). *Co-Design of Innovations with Teachers: Definition and Dynamics*.

Salac, J., Landesman, R., Druga, S., & Ko, A. J. (2023). Scaffolding Children's Sensemaking around Algorithmic Fairness. *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, 137–149. https://doi.org/10.1145/3585088.3589379

Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019). Envisioning AI for K-12: What Should Every Child Know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), Article 01. https://doi.org/10.1609/aaai.v33i01.33019795

Walsh, B., Dalton, B., Forsyth, S., & Yeh, T. (2023). Literacy and STEM Teachers Adapt AI Ethics Curriculum. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(13), 16048–16055. https://doi.org/10.1609/aaai.v37i13.26906

Williams, R., Ali, S., Devasia, N., DiPaola, D., Hong, J., Kaputsos, S. P., Jordan, B., & Breazeal, C. (2023). AI + Ethics Curricula for Middle School Youth: Lessons Learned from Three Project-Based Curricula. *International Journal of Artificial Intelligence in Education*, *33*(2), 325–383. https://doi.org/10.1007/s40593-022-00298-y

Yue, M., Jong, M. S.-Y., Dai, Y., & Lau, W. W.-F. (2025). Students as AI literate designers: A pedagogical framework for learning and teaching AI literacy in elementary education. *Journal of Research on Technology in Education*, 1–22. https://doi.org/10.1080/15391523.2025.2449942

Acknowledgement

This work was funded by the National Science Foundation, under awards DRL-2049029 and DRL-2048502.