

Measuring Science Teachers' Generative AI Literacy for Classroom Assessment: Performance Assessment

Ruiping Huang and Yue Yin

College of Education, University of Illinois Chicago
1040 W Harrison St, Chicago, IL, 60607, USA

Abstract

We report preliminary analyses from an ongoing study using performance-based tasks to assess science teachers' generative AI (GenAI) literacy for classroom assessment and compare these scores with self-reported GenAI use and confidence. Preliminary findings indicate moderate correlation with use frequency but weak correlation with confidence, highlighting performance assessments' value.

Keywords: Generative AI; AI literacy; classroom assessment; performance assessment

1 Background

Widely available generative AI (GenAI) tools such as ChatGPT and Gemini have demonstrated substantial potential to strengthen classroom assessment, a critical component of instruction that provides evidence on students' strengths, needs, and learning progress to inform teachers' instructional decisions (Baidoo-Anu et al., 2025; Risko and Walker-Dalhouse, 2010). From item generation and automated scoring to personalized feedback and support for academic integrity (Naseer et al., 2024; Alier et al., 2024), GenAI is transforming how teachers design, administer, and interpret assessments.

However, teachers face substantial challenges in implementing GenAI, as evidenced by their limited knowledge, low confidence and interest, negative attitudes, ethical concerns, disruptions to traditional assessment practices, and difficulties in interpreting GenAI outputs (Yan et al., 2024; Alier et al., 2024; Nah et al., 2023; Celik et al., 2022; Zhai and Nehm, 2023). Effectively addressing these demands requires cultivating teachers' GenAI literacy (Yan et al., 2024; Nyaaba, 2024), particularly for classroom assessment.

It is critical to measure teachers' GenAI literacy specifically for classroom assessment using

mixed-methods approaches for the following reasons. First, most existing AI literacy frameworks focus on students rather than teachers (Zhou et al., 2025; Lintner, 2024). Second, current assessments rely primarily on self-report surveys (Knoth et al., 2024; Ding et al., 2024). Very few studies used performance assessment, that is, how teachers use GenAI in teaching practice. They do not fully capture GenAI-specific competencies, such as prompt design, output evaluation, and ethical reasoning (Zhang and Magerko, 2025; Gu and Ericson, 2025). Third, most studies address general digital or AI literacy and are not tailored to the pedagogical and assessment demands of day-to-day classroom practice (Sparks et al., 2024).

2 Aims

To address these gaps, we developed performance assessment tasks to measure middle and high school science teachers' GenAI literacy for classroom assessment. Our study is designed to answer the following questions:

- (a) How reliable and valid are the performance assessment tasks?
- (b) How is the science teachers' GenAI literacy measured by the performance assessment tasks?
- (c) How is the science teachers' GenAI literacy measured by the performance assessment tasks compared with that measured by the self-reported survey?

3 Samples

We are currently collecting data. Twenty-one in-service middle and high school science teachers from U.S. Midwest school districts have been recruited and completed a self-reported survey. Table 1 presents the participants' characteristics and descriptive statistics.

Table 1: Participant Characteristics and Descriptive Statistics (N = 21)

Variable	Value	n (%)	M (SD)
Gender	Female	11 (52.38)	
	Male	10 (47.62)	
Race/Ethnicity	White	18 (85.71)	
	Hispanic/Latino	1 (4.76)	
	Asian	1 (4.76)	
	Other: Mexican American	1 (4.76)	
Education Level	Bachelor's degree	8 (38.10)	
	Master's degree	11 (52.38)	
	Doctoral degree	2 (9.52)	
Age Group	20–29	2 (9.52)	
	30–39	13 (61.90)	
	40–49	5 (23.81)	
	50–59	1 (4.76)	
School Level	Middle school	11 (52.38)	
	High school	10 (47.62)	
Subjects (multiple choices)	General Science	12 (57.14)	
	Biology	3 (14.29)	
	Chemistry	7 (33.33)	
	Physics	1 (4.76)	
	Earth/Space Science	3 (14.29)	
Student Access to Devices	Yes	21 (100)	
Teaching Experience (years)			11.29 (5.53)
Student IEP/504 (%)			21.00 (0.09)
English Learners (%)			20.00 (0.26)
Prior AI PD Participation hours			3.90 (8.55)
Self-Directed AI Learning Hours			5.57 (7.32)

Note. Categories not selected by any participant (e.g., Black, Pacific Islander) are omitted for clarity. The complete survey was in Appendix A.

Among them, nine teachers have completed the performance assessment tasks. We anticipate 30 science teachers will have completed both the self-reported survey and performance assessment tasks by the end of August 2026.

4 Methods

To guide the design of the instruments, we developed a conceptual framework (Figure 1) grounded in the iterative classroom assessment cycle (Bonner and Chen, 2019), which includes four phases: developing assessments, collecting responses, interpreting responses, and using results (Table 2). This cycle builds on foundational assessment literature (Buhagiar, 2006; Gottlieb and Katz, 2020; Ruiz-Primo and Furtak, 2007; Hajj-Hassan et al., 2014) and is refined to capture how teachers can engage with GenAI tools at each phase (Sipahioglu, 2024; Chaivisit et al., 2024; MacDowell et al., 2024).

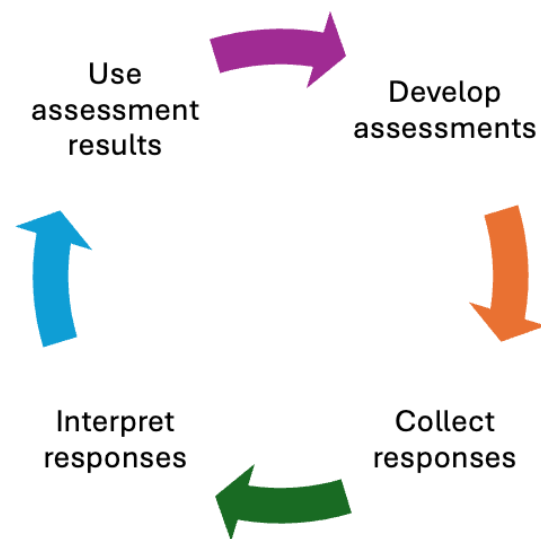


Figure 1: Teachers' classroom assessment cycle

At every phase of classroom assessment, GenAI can play a supportive role (Table 2). For example, in the assessment-development phase, GenAI can draft assessment items and tasks; in the response-collection phase, it can increase accessibility by adapting content and diversifying formats (e.g., visual or audio aids, translation); in the response-interpretation phase, it can assist with automated grading and organizing student work; and in the

use-of-results phase, it can suggest follow-up assessments and learning resources. We used this alignment between GenAI functions and assessment phases to develop instruments to measure teachers’ GenAI literacy, specifically within the context of classroom assessment.

Table 2: GenAI-Supported Functions Across Phases of Teachers’ Classroom Assessment

Phase	How GenAI Can Help Teachers for Classroom Assessment
Develop assessments	-Draft assessments (e.g. items, performance tasks) -Align assessments to standards/rubrics -Differentiate assessment materials by level or learner profile
Collect responses	-Make assessment inclusive and accessible by adapting content, diversifying formats, and reducing bias. -Simulate student responses to test clarity and difficulty -Transcribe/Translate student input when needed
Interpret responses	-Process and organize student responses (e.g., summarizing, tagging, identifying incomplete data) -Assist with automated grading
Use assessment results	-Draft feedback summaries and insights -Draft post-assessment reports, and action plans -Recommend follow-up assessments or learning resources

Appendix A is the performance assessment tasks we developed. Appendix B is the scoring rubric we developed to study teachers’ GenAI literacy. Appendix C is the self-reported survey measuring GenAI literacy.

5 Anticipated Results

5.1 RQ1. Psychometric Features

We are still collecting data, refining the final performance assessment rubrics, and training coders. Table 3 shows the preliminary scoring results of the nine teachers. We found that the internal consistency of the four items was .78. Because the scoring rubrics and coder training are still being finalized, inter-rater reliability has not yet been established. We will report inter-rater reliability evidence in the full paper after rubric refinement and coder training are completed. The internal consistency of the confidence measure for the four phases are 0.89, 0.93, 0.95, and 0.95 respectively. The internal consistency of the frequency measure of the four phases are 0.86, 0.81, 0.93, and 0.79 respectively.

Table 3: Descriptive Statistics of Performance Assessment Scores (N = 9)

	Minimum	Maximum	Mean	Std. Deviation
Prompt	1.00	2.00	1.78	.44
Format and File Use	.00	2.00	1.00	.87
Communication	.00	2.00	.89	.93
Critical Evaluation	.00	2.00	1.00	.71
Total	.50	2.00	1.17	.57

5.2 RQ2. GenAI Literacy Profile

Preliminary analyses indicate the following patterns:

First, teachers vary in their proficiency with GenAI for classroom assessments on their performance assessment tasks (Table 3). As shown in Appendix D, teacher 6’s prompt includes context, content, and output specifications, and instructional clarity, and scored 2, while teacher 1 includes only context and content specifications and scored 1 for the prompt.

Second, teachers’ performance differs significantly across the four dimensions (Figure 2), $F(3, 24) = 4.59$, $p = 0.011$. Post hoc comparisons in a repeated ANOVA with the Sidak adjustment show that teachers, on average, scored significantly higher than the critical evaluation ($p = .004$).

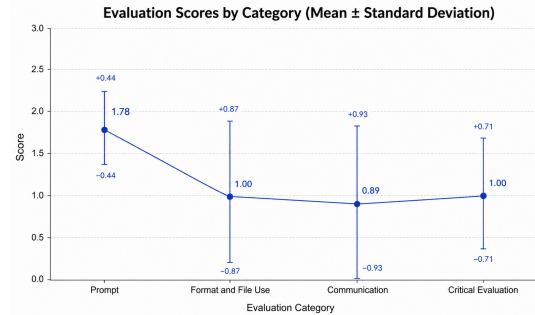


Figure 2: Scoring of the four dimensions

Third, we plan to further score and compare teachers’ performance across dimensions to examine additional validity evidence. For example, teachers may score relatively high on one dimension, such as format and file use, but lower on another dimension, such as critical evaluation or communication. Such patterns can help identify the aspects of GenAI literacy that may require additional instructional or professional development support.

5.3 RQ3. Relationship with Self-reported Scales

Preliminary results show that teachers’ performance scores are moderately correlated with their

frequency of GenAI use, $r = .63$, $p = .07$; but the correlation between their self-rated confidence and performance score is relatively weak, $r = .25$, $p = .52$.

We will further compare teachers' scores on the performance assessment tasks across four assessment phases and their self-reported frequency of GenAI use and GenAI confidence.

6 Anticipated Conclusions

Our study will fill the following gaps: First, instruments are needed to measure teachers' AI literacy for classroom assessment. Second, performance assessments are lacking to measure AI literacy, while most scales are self-reported surveys.

Our performance assessment tasks will help better understand teachers' current use of GenAI in their assessment practice. Although the current performance assessment tasks were designed for middle and high school science teachers, the classroom assessment framework and scoring rubrics are general rather than subject-specific. Thus, the tasks can be adapted for teachers in other subject areas or educational contexts by revising the classroom scenarios, disciplinary content, and expected assessment products.

The comparison between teachers' performance assessment and their self-reported data will help inform the validity study of measurements.

These findings should be interpreted as preliminary. Data collection is ongoing, and the full paper will provide an updated sample description, report inter-rater reliability after rubric refinement and coder training, and examine additional validity evidence through relationships among the performance dimensions as the available data allow.

References

- Marc Alier, Francisco José García Peñalvo, and Jorge D. Camba. 2024. [Generative Artificial Intelligence in Education: From Deceptive to Disruptive](#). *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(5):5–14.
- David Baidoo-Anu, Amirhossein Rasooli, Christopher DeLuca, and Liying Cheng. 2025. [Conceptions of classroom assessment and approaches to grading: teachers' and students' perspectives](#). *Education Inquiry*, 16(3):411–439.
- Sarah M. Bonner and Peggy P. Chen. 2019. *Systematic Classroom Assessment: An Approach for Learning and Self-Regulation*, 1 edition. Routledge.
- Michael A. Buhagiar. 2006. [The classroom assessment cycle within the alternative assessment paradigm: exploring the role of the teacher](#).
- Ismail Celik, Muhterem Dindar, Hanni Muukkonen, and Sanna Järvelä. 2022. [The Promises and Challenges of Artificial Intelligence for Teachers: a Systematic Review of Research](#). *TechTrends*, 66(4):616–630.
- Sarinporn Chaivisit, Tataleni Iita Asino, Sukanda Jongsermtrakoon, Penny Thompson, Fatemeh Rezaie, and Sutitthep Siripipattanakul. 2024. [Empowering Educators With Generative AI Tools and Support](#). In *Advances in Educational Technologies and Instructional Design*, pages 56–81. IGI Global.
- Lu Ding, Sohee Kim, and R. Allan Allday. 2024. [Development of an AI literacy assessment for non-technical individuals: What do teachers know?](#) *Contemporary Educational Technology*, 16(3):ep512.
- Margo Gottlieb and Anne Katz. 2020. [Assessment in the Classroom](#). In *The Encyclopedia of Applied Linguistics*, pages 1–8. Wiley.
- Xingjian Gu and Barbara J. Ericson. 2025. [AI Literacy in K-12 and Higher Education in the Wake of Generative AI: An Integrative Review](#). *arXiv preprint*.
- Mohamad Hajj-Hassan, Bassam Hussein, Khaled Chahine, and Amin Haj-Ali. 2014. [A Sustainable Approach for Program Students Outcomes Assessment: The Case of the Department of Biomedical Engineering at the Lebanese International University](#). *Journal of Modern Education Review*, 4(7):555–563.
- Nils Knoth, Marie Decker, Matthias Carl Laupichler, Marc Pinski, Nils Buchholtz, Katharina Bata, and Ben Schultz. 2024. [Developing a holistic AI literacy assessment matrix – Bridging generic, domain-specific, and ethical competencies](#). *Computers and Education Open*, 6:100177.
- Tomáš Lintner. 2024. [A systematic review of AI literacy scales](#). *npj Science of Learning*, 9(1):50.
- Paula MacDowell, Kristin Moskalyk, Katrina Korchinski, and Dirk Morrison. 2024. [Preparing Educators to Teach and Create With Generative Artificial Intelligence](#). *Canadian Journal of Learning and Technology*, 50(4):1–23.
- Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. [Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration](#). *Journal of Information Technology Case and Application Research*, 25(3):277–304.
- Fawad Naseer, Muhammad Usama Khalid, Nafees Ayub, Akhtar Rasool, Tehseen Abbas, and Muhammad Waleed Afzal. 2024. [Automated assessment and feedback in higher education using generative AI](#). In *Transforming education with generative AI*.

- Prompt engineering and synthetic content creation*, pages 433–461. IGI Global Scientific Publishing.
- Matthew Nyaaba. 2024. [Transforming Teacher Education in Developing Countries: The Role of Generative AI in Bridging Theory and Practice](#). *arXiv preprint*.
- Victoria J. Risko and Doris Walker-Dalhouse. 2010. [Making the Most of Assessments to Inform Instruction](#). *The Reading Teacher*, 63(5):420–422.
- Maria Araceli Ruiz-Primo and Erin Marie Furtak. 2007. [Exploring teachers’ informal formative assessment practices and students’ understanding in the context of scientific inquiry](#). *Journal of Research in Science Teaching*, 44(1):57–84.
- Mete Sipahioglu. 2024. [Empowering Teachers With Generative AI Tools and Support](#). In *Advances in Educational Technologies and Instructional Design*, pages 214–238. IGI Global.
- Jesse R. Sparks, Teresa M. Ober, Caitlin Tenison, Burcu Arslan, Ido Roll, Paul Deane, Diego Zapata Rivera, Reginald M. Gooch, and Tenaha O’Reilly. 2024. [Opportunities and challenges for assessing digital and ai literacies](#). Technical report, ETS Research Institute.
- Lixiang Yan, Samuel Greiff, Ziwen Teuber, and Dragan Gašević. 2024. [Promises and challenges of generative artificial intelligence for human learning](#). *Nature Human Behaviour*, 8(10):1839–1850.
- Xiaoming Zhai and Ross H. Nehm. 2023. [AI and formative assessment: The train has left the station](#). *Journal of Research in Science Teaching*, 60(6):1390–1398.
- Chengzhi Zhang and Brian Magerko. 2025. [Generative AI Literacy: A Comprehensive Framework for Literacy and Responsible Use](#). *arXiv preprint*.
- Xinyan Zhou, Yan Li, Ching Sing Chai, and Thomas K. F. Chiu. 2025. [Defining, enhancing, and assessing artificial intelligence literacy and competency in K-12 education from a systematic review](#). *Interactive Learning Environments*, 33(10):5766–5788.

Appendix A. Performance Assessment Tasks

Please use the ChatGPT account we created for you to complete the following tasks.

Website: <https://chatgpt.com/>

Account: xxx

Password: xxx

Task 1: Develop Assessments

- You are preparing an assessment at the end of a science unit. Please use ChatGPT to generate a classroom assessment to measure students' learning outcome and its corresponding solutions.

Task 2: Collect responses

- Please use ChatGPT to create sample students' responses to an existing science assessment from your science class.

Task 3: Interpret Responses

- Suppose the following students are your real students in your classroom, below are their responses to a science question (comparing types of natural sugars). Use ChatGPT to help you score/analyze the students' responses.

Assessment Tasks for Students: Comparing Types of Natural Sugars

Amy wants to know if the sugar found in honey, milk, sugarcane, and apples are the same. To find this out, she chemically removed the sugar from each food and recorded the properties of the sugar in Table 1.

Table 1: Characteristic properties of sugar in different foods

Source of sugar sample	Density	Solubility in Water	Melting Point
Honey 	1.69 g/cm ³	Yes	103 °C
Milk 	1.53 g/cm ³	Yes	202 °C
Sugarcane 	1.59 g/cm ³	Yes	186 °C
Apple 	1.70 g/cm ³	Yes	103 °C

Which of the above sources has the same type of sugar? Why?

Right Answer:

Apple and honey have the same type of sugar, as they share the same density (with any small differences likely due to measurement error) and melting point. Both density and melting point are characteristic properties used to identify substances.

Here are 6 students' responses:

Alex Martinez: Well the one that has the similar sugar is honey and apple because there only 1 percent from each other. Honey and apple, because there like 1 sugar away from each other. And milk and sugar are 6 percent of sugar away from each other.
Emily Chen: My conclusion is that none have the same sugar and two of them have the same melting point and they all have the same solubility in water. Differences are density also sugar can and milk are different in melting point from honey and apple. Honey and apple have the same melting point. All of them (honey,milk,sugar cane, apple) have the same solubility in water. They all have a different density. Milk's and sugar cane's melting point are different from honey's and apple's melting point.
Jordan Williams: A lot of foods have the same sugar but some foods have a decent amount of sugar the differences is that sugar can be in anything the simularites is that they have foods that are the same with sugar The similarities of the foods is that some of the food have more sugar than the other kinds of food and some people put the same sugar in foods and put water in them for the differences
Sophia Nguyen: What i can conclude from the pattern is that the apple and honey sugar are almost the most because they have the same melting point and they almost have the amount of density. The similarities or differences i would look to tell weather they have the same sugar is the Density and the melting point.
Michael Johnson: I can draw a conclusion that the sugar of the sources from both the honey and the apple are the same because their characteristic properties are either the same or very, very similar. For example, the density of the honey is 1.69 g/cm cubed, and the apple's density is 1.70g/cm cubed, so the difference between both the densities is only 0.01g/cm cubed. Additionally, their are both soluble and have the melting points, which is 103 degrees Celsius. However, the only thing similar between the milk and the sugar cane is that they are both soluble, and other than that, the density of each are not the same, and they have two, very different melting points, meaning that they are most likely not made of the same sugar. In order to tell whether or not any of the foods contained the same sugar, I would look to see if the density and melting point were the same or, at least, similar. Also, I would make sure that the solubility of the sugar from the different sources are the same. In conclusion, I would make sure that the characteristic properties between the sugar from the multiple sources were the same because the properties are what make up a substance and what makes each one similar or different, so if all of the properties for each source are the same, then it would mean that the sugar used in each sample was also the same.
Kayla Thompson: Claim: honey and apple have the same sugar. evidence: the density (1.69 and 1/70 g/cm ³) and melting point are the same (103 C). Reasoning: unique substances that have unique characteristic properties.

Task 4: Use Assessment Results

- Building on the analysis results in Task 3, please use ChatGPT to support teaching and learning based on these assessment results.

Appendix B. Performance Assessment Scoring Rubrics

Criterion	Definition (What is being evaluated)	0 = Novice	1 = Emerging	2 = Experienced
Prompt Construction	How clearly and effectively the teacher constructs prompts to guide GenAI in achieving the intended task goals. Prompts are evaluated based on the presence of four core prompt features: (a) Context (e.g. grade level, student characteristics); (b) Content (such as plate tectonics as topic or content) (c) Output (Question format, structure, length, style, or inclusion of examples); (d) Instructional clarity (clear, complete, and actionable instruction, explicit guidance such as step-by-step directions, role assignment, or explicit constraints or criteria, such as align with Indiana standard)	Demonstrates none of the core prompt features. Prompts largely copy or restate the original task instructions with little or no modification.	Demonstrates 1–2 of the core prompt features across the four tasks (features may appear separately across prompts or in combination within a single prompt).	Demonstrates 3–4 core prompt features across the four tasks (features may appear separately across prompts or in combination within a single prompt). Prompts are aligned with task goals.
Format and File Use	The extent to which the teacher can appropriately use different input and output formats with GenAI to support task completion, including requesting responses in useful forms and working with uploaded materials.	Uses GenAI primarily in plain text form. Does not request specific output formats (e.g., tables, worksheets, doc, pdf, images, canvas-based drafts) and does not use file upload features.	Uses 1–2 distinct formats or file interactions (e.g., requests a table or worksheet, or uploads a PDF/document/image) or attempts format use 1–2 times.	Purposefully selects and uses appropriate formats to support the task (e.g., asks GenAI to organize content into tables, worksheets, structured documents, or canvas-based drafts) and effectively works with uploaded materials when relevant (e.g., PDFs, images). Demonstrates ≥ 3 instances of format or file use.
Interactive Communication (B, C)	The extent to which the teacher engages in ongoing, responsive communication with GenAI during task completion, using follow-up interactions to clarify, refine, or redirect responses.	Accepts initial GenAI outputs with no follow-up interaction or clarification.	Engages in 1–2 follow-up interactions (e.g., minor revisions or clarifications), but these interactions do not improve subsequent GenAI responses.	Actively communicates with GenAI through ≥ 3 purposeful follow-up requests, clarifications, or redirections informed by prior outputs.
Critical Evaluation and Decision-Making (C, E) (Sullivan, Michael, et al., 2024)	The extent to which the teacher critically evaluates GenAI outputs (e.g., accuracy, appropriateness) and uses professional judgment to accept, revise, or reject outputs to make task-relevant decisions.	Accepts outputs without verification or critique.	Identifies issues in the GenAI output (e.g., inappropriate content, unclear format) but conducts limited evaluation or doesn't follow up or give up easily.	Not only identifies issues, but also actively verifies or challenges outputs (e.g., checks sources, requests evidence, or asks GenAI to justify or revise responses).

Note: B = Behavioral; C = Cognitive; E = Ethical

Appendix C. Self-reported Scales to Measure Teacher's GenAI Literacy

Phase 1: Develop Assessments

- How frequently did you use AI for each of the following assessment activities in spring 2025? (Left)
- How confident are you now in using AI for each of the following assessment activities? (Right)

	My Frequency in Spring 2025					My Current Confidence Level				
	Never (1)	Rarely (once a semester) (2)	Sometimes (several times a semester) (3)	Often (monthly) (4)	Always (weekly) (5)	Strongly unconfident (1)	Unconfident (2)	Neutral (3)	Confident (4)	Strongly confident (5)
Use AI to modify existing classroom assessments.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Use AI to draft classroom assessments (e.g. items and performance-based tasks).	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Use AI to develop scoring rubrics or standards for classroom assessments.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Use AI to develop solutions to answers for classroom assessments.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Use AI to design classroom assessments that require students to engage in their own thinking instead of relying solely on AI.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Phase 2: Collect Responses

- How frequently did you use AI for each of the following assessment activities in spring 2025? (Left)
- How confident are you now in using AI for each of the following assessment activities? (Right)

	My Frequency in Spring 2025					My Current Confidence Level				
	Never (1)	Rarely (once a semester) (2)	Sometimes (several times a semester) (3)	Often (monthly) (4)	Always (weekly) (5)	Strongly unconfident (1)	Unconfident (2)	Neutral (3)	Confident (4)	Strongly confident (5)
Use AI to translate and administer classroom assessment to accommodate English learners.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Use AI to modify and administer classroom assessments in accessible formats for learners with special needs (e.g., audio and visual versions).	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Use AI to revise and administer classroom assessment materials to accommodate students with different culture backgrounds.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Use AI to simulate student responses to test the clarity and difficulty of classroom assessments.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Use AI to transcribe	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

student responses when needed. Use AI to translate student responses when needed.

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Q20 Phase 3: Interpret Responses

- How frequently did you use AI for each of the following assessment activities in spring 2025? (Left)
- How confident are you now in using AI for each of the following assessment activities? (Right)

	My Frequency in Spring 2025					My Current Confidence Level				
	Never (1)	Rarely (once a semester) (2)	Sometimes (several times a semester) (3)	Often (monthly) (4)	Always (weekly) (5)	Strongly unconfident (1)	Unconfident (2)	Neutral (3)	Confident (4)	Strongly confident (5)
Use AI to summarize student responses to classroom assessments.	○	○	○	○	○	○	○	○	○	○
Use AI to score students' responses to objective classroom assessments (e.g., multiple-choice, true/false, matching).	○	○	○	○	○	○	○	○	○	○
Use AI to score students' responses to open-ended classroom assessments.	○	○	○	○	○	○	○	○	○	○
Use AI to analyze students' responses to classroom assessments.	○	○	○	○	○	○	○	○	○	○
Use AI to group students based on their responses to classroom assessments.	○	○	○	○	○	○	○	○ ○		○

Phase 4: Use Assessment Results

- How frequently did you use AI for each of the following assessment activities in spring 2025? (Left)
- How confident are you now in using AI for each of the following assessment activities? (Right)

	My Frequency in Spring 2025					My Current Confidence Level				
	Never (1)	Rarely (once a semester) (2)	Sometimes (several times a semester) (3)	Often (monthly) (4)	Always (weekly) (5)	Strongly unconfident (1)	Unconfident (2)	Neutral (3)	Confident (4)	Strongly confident (5)
Use AI to draft feedback for students based on classroom assessment results.	○	○	○	○	○	○	○	○	○	○
Use AI to recommend follow-up assessments or learning resources based on classroom assessment results.	○	○	○	○	○	○	○	○	○	○
Use AI to reflect on student learning progress using classroom assessment results.	○	○	○	○	○	○	○	○	○	○
Use AI to help make instructional decisions based on classroom assessment results to assignments.	○	○	○	○	○	○	○	○	○	○
Use AI to support my communication with parents on classroom assessment results.	○	○	○	○	○	○	○	○	○	○

Appendix D. Examples of Teachers' Prompts in Phase 1. Assessment Generation

- Teacher 6: “*You are an experienced chemistry teacher with a masters degree in chemistry. You are teaching an honors chemistry class for advanced sophomores that will get dual credit at completion of the course through a community college [context specification]. I would like you to make an end of unit assessment to measure student learning on kinetics and equilibrium [content specification] that is between 20 and 25 questions that must include questions about an ice table, questions on le chatleier principle, and a problem where students must determine rate order from lab data [output specification]. I would like you to also create a key for the assessment. The assessment must fit within indiana state standards in chemistry and A typical Chemistry 101 and 102 course at the university level [Instructional clarity].*”
- Teacher 1: “*Generate an end-of-unit assessment testing high school students' [context specification] knowledge of plate tectonics, specifically the three main types of plate boundaries as well as intraplate activities [content specification].*”