A Initial Assessment Design

The initial assessment consists of six multiple-choice questions. Each question presents three answer choices, with only one being correct.

A.1 Learning Objectives (LOs) and AI4K12 Standard

There are 3 initial learning objectives (as shown in Figure 1 left column) that are derived from AI4K12 guideline [?] and teachers' real needs [?] for instructional materials to teach students how to use AI for learning. There are two assessment questions and corresponding instructional activities in the learning session under each LO.

A.2 Multiple-Choice Questions as a Low-cost Scalable Assessment Method

Multiple Choice Questions (MCQs) enable automatic grading while providing consistent evaluations across diverse learner groups. When carefully crafted, MCQs can target any cognitive level in Bloom's Taxonomy [?], making them a versatile and effective component of our assessment approach. Therefore, we chose MCQs as the only assessment question type in the initial version.

A.3 Applying scenario-based analytical assessments to measure students' skill level

Scenario-based assessments have been applied to measuring skills in varied domains, such as writing [?], mathematics [?] and science [?]. This situated assessment experience can engage learners and facilitate deep thinking by presenting realistic, contextual problems that promote critical thinking and knowledge application [? ?]. Furthermore, as our instructional goals were competency-based, scenario-based assessments that apply knowledge within specific scenarios are suitable to test students' abilities. Therefore, we situated the MCQs into different scenarios based on the corresponding learning objectives.

B Iterated Assessment Design

The iterated assessment consists of 15 questions in total, comprising ten True or False (TF) questions and five Open-Ended (OE) questions. TF1 (first TF question) to TF6 and OE1 is *LO-AI Capacity*; TF7 to TF10 and OE2, OE3 are under *LO-When Learning Happens*; OE4 and OE5 are under *LO-Prompt Writing*.

B.1 Iteration Rationale

Changes have been made to the learning objectives (*LOs*), assessment question types (*MCQ*-multiple choice question; *TF*-true or false question; *OE*-open-ended question), and question content. The iterated version of assessment questions can be found in Figure 1-right.

B.1.1 Refocus and reorganize the assessment's targeted learning objectives. Firstly, as questions under *LO-Effective Question* (*Formulate effective questions for AI chatbots*) assess a subset of knowledge in *LO-Prompt Writing* with more fine-grained granularity, we only kept *LO-Prompt Writing* in the iterated assessment. In addition, students' pre-test scores exhibited a ceiling effect on questions under **LO-When to Use AI** (*Determine when using AI is good/bad for learning*), and we hypothesized this is because the previous questions focused only on avoiding the use of AI to generate direct answers. More challenging scenarios, where it's difficult to decide and when learning happens need to be considered, and should also be covered. Hence we refocus the learning objective with an emphasis on when learning happens.

Also, while the original assessment did not explicitly include LOs and questions on AI's capabilities, many students reported AI Capabilities (e.g. AI can create practice questions) as their newly gained knowledge in the post-test survey. This finding highlights both the learners' lack of prior knowledge about AI's capabilities and the effectiveness of the instructional materials in conveying this concept. As a result, we added **LO-AI Capabilities** (*Identify capabilities of AI chatbot to help you learn*) to the assessment.

B.1.2 Replace multiple-choice questions with True / False questions and open-ended questions. The lack of information about learners' prior knowledge can reduce the effectiveness of MCQ distractors. Students may compare options and select the correct answer through elimination or guessing. In contrast, TF questions require students to make a decision on each item individually, offering more data points to assess both the learners' true knowledge level and the quality of the question items, particularly at lower granularity [?].

In addition, to overcome difficulties in forming high-quality distractors in MCQs, one approach is to collect learners' common misconceptions as MCQ items. OE1, OE2, and OE3 are designed as a misconception collectors for **LO-AI Capability** and **LO2-When Learning Happens** to facilitate further iterations on assessment questions.

Although the ceiling effect in the MCQ scores for **LO-Prompt Writing** suggested a high level of prior knowledge on this learning objective, the students' performance in prompt writing activities during the learning session indicates otherwise. This discrepancy may be due to the fact that unlike LOs in **Design Rationale 2.5**, **LO-Prompt Writing** involves higher-order cognitive skills (falling under the *analyze* and *create* levels of Bloom's Revised Taxonomy [?]), and developing multiple-choice questions that effectively assess these higher cognitive levels is challenging even for expert instructors in traditional disciplines [?], let alone in an emerging field like AI Literacy. Therefore, we designed OE4 (*(Given a prompt and the AI Chatbot response) Do you think the question can generate good learning material for your quiz preparation? Why or why not?*) to assess students' analyzing skill, and OE5 (*Rewrite the question to AI to generate better preparation questions*) to assess prompt writing skill.

B.1.3 Add abstract-level questions to increase the variation of question difficulties. To more accurately estimate learners' prior knowledge, assessment questions should be designed with varying levels of difficulty and knowledge depth under the same learning objective. As such, abstract-level questions are included alongside concrete, scenario-based questions to assess learners across a broader spectrum of difficulty within the same objective.
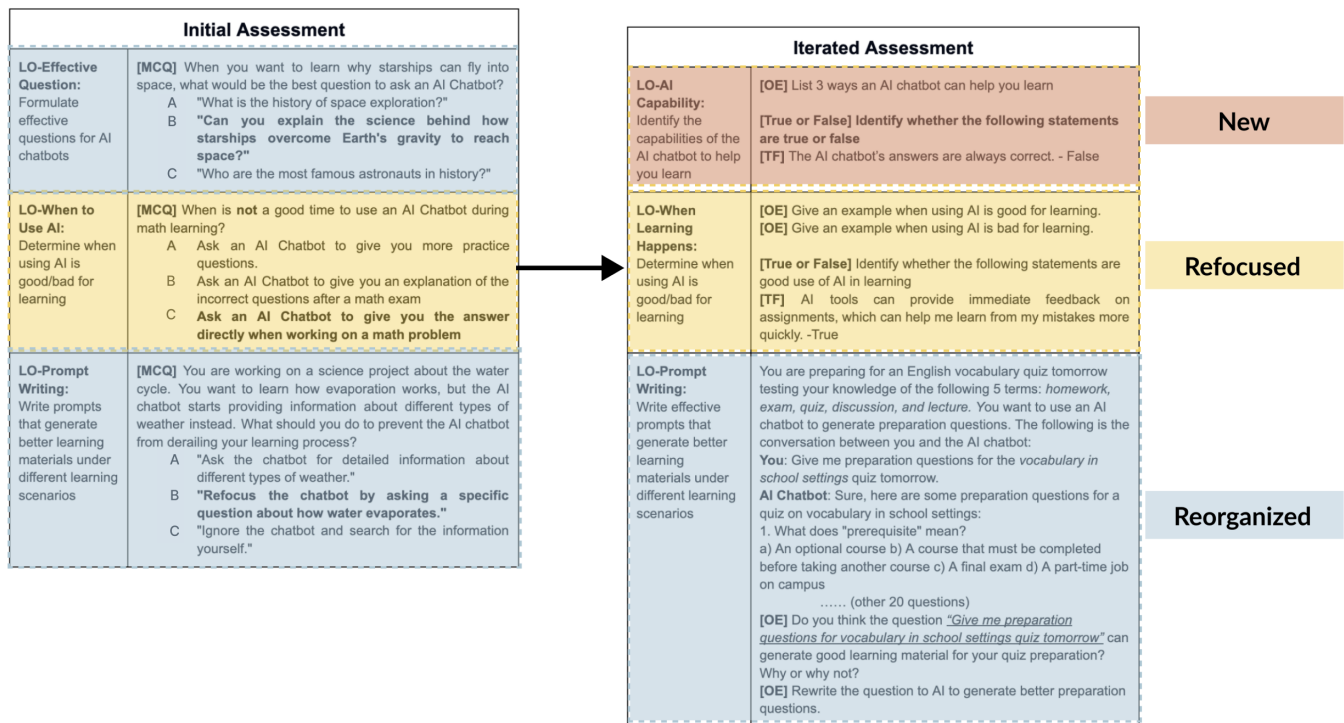
Figure 1: Initial and Iterated Assessment Design