Developing an AI-Based Mathematics Tutor: A Type 1 Design and Development Research Study for a Scaffolded LLM Tutoring Design Framework

Andrew Egbert

Instructional Design and Technology

EDLT 7733: Design and Development Research

Dr. John H. Curry III

March 18, 2026

Abstract

This research is a formative validation of the initial deployed version of the Scaffolded LLM Tutoring Design Framework (SLTDF), targeting community college students preparing for math placement testing. Prior research in intelligent tutoring systems and learning science demonstrates that mastery-gated tutoring guided by calibrated feedback provides measurable gains. ITS development requires specialized infrastructure and technical knowledge impractical for individual designers. LLM-based tutoring is an emergent technology that lacks a structured framework connecting prompt design to learning theory, making it difficult to reliably design and implement. The Math Tutor GPT is the latest of eight development iterations. Eight expert reviewers evaluated through seven scenario-based interactions. This formative expert review (n=8) used purposive maximum-variation sampling across AI/LLM prompting, mathematics education, and instructional design domains. Six of seven prompted tutoring behaviors achieved moderate-to-high design fidelity (grand mean = 3.02), providing initial formative validation of SLTDF as a transferrable framework for prompt-engineered LLM tutoring.

Keywords: design and development research, large language models, math Tutor, remedial mathematics, scaffolding, formative evaluation, intelligent tutoring systems, community college, placement preparation

Table of Contents

Right-click and choose Update Field (or press F9) to populate the Table of Contents.

Introduction

Background and Context

To properly encapsulate the purpose of this design and development research project, the curious reader will wish to turn to the earliest pages in my education. As it happens, I was not always garrulous among my classmates. Often, an erstwhile glance might discover me silently inhabiting my own world while my peers transformed the classroom into a boisterous chamber. Of course, such a scene poses little issue when the teacher solicits the class for answers, especially during a mathematics class.

This situation was not entirely conducive to the kind of in-person interaction that required learners to venture to the front desk to ask for assistance. As it happened, I still learned the content, mostly, but a mental schema regarding the structure of learning and the vagaries best suited to it took root. It was not an instantaneous realization, but it became apparent that the K-12 educational system was focused firmly on serving the bulk of the personality spectrum. Those outside this set would often find themselves left behind or forced to make the best of being uncommon, learning to blend in wherever necessary.

It was some time after this, during my last year of high school, that I took my first true correspondence course in accounting. Most will agree that rectifying columns of numbers is tedious, unimaginative work at best, which was certainly the case. However, the epiphany that I'd been heading toward began to further coagulate in my mind. I was no longer in a classroom, and I did not need to inhabit others' space to succeed. It was a sublime, eye-opening experience in a world where grids of desks formed the foundation for traditional classroom learning.

It appears I have built a reputation for developing technical courseware, software solutions, and innovative products at the College of Western Idaho. Having extensively experimented with complex and highly effective instruction sets for Tutoring bots, including standalone general pretrained transformers (GPT) for math students, Jennifer Miller, a professor at the Math Solutions Center, reached out to ask me to create an AI bot that could help remedial students practice for math placement tests.

Fascinated by the opportunity that had landed in my lap, I developed the product and wove seven educational theories into its code, moving from theoretical to practical. These ideas had to work in practice, not simply a thought experiment. The latest version, v4.3, would be scrutinized by expert reviewers. It appears that the product demonstrates many of the enhanced cognitive interaction that should enable it to perform well with learners.

A reader might imagine from the preceding narrative that I hated the classroom, but this is not the case. In truth, I had no actual problem with it, but, having read a considerable amount of science fiction, I dreamed of the day when artificial intelligence would finally become a reality. How might learning change with a tool unlike any other, that thinks for itself and has been trained on most of the knowledge accrued by humankind?

Problem Statement

Established tutoring methods and strategies are common in collegiate environments. However, the translation of these methods to a Large Language Model (LLM) AI-based intelligent Tutoring system (ITS) requires the designer to convert learning principles that result in effective transfer enabled by consistent and controllable

Tutoring behaviors. While General Pretrained Transformer (GPT) output naturally differs between sessions, the ability to maintain overall consistency requires studied technical design knowledge, though this ability only represents part of the task. Continued research is required to determine whether the LLM Tutor actually produces measurable improvements in placement-aligned learning outcomes. Only a thorough analysis of the interaction patterns will determine whether the desired results occur.

Data Collection

This study collected reviewer ratings and open-ended recommendations from eight expert reviewers via a Qualtrics survey hosted under Idaho State University's institutional license. The instrument evaluated the Math Tutor GPT (Custom GPT V4.1) against the seven evidence-informed design features specified by the SLTDF framework, producing data that addressed both research questions: design fidelity (RQ1) and recommended refinements (RQ2).

Researcher Qualifications

The researcher has worked with large language models since their introduction in 2022, beginning with the ChatGPT Research Preview in November of that year, and is currently in the top 1% of all users worldwide. He holds a master's degree in education and a graduate certificate in instructional design and works as a professional instructional designer at the College of Western Idaho. Additionally, he is pursuing a doctorate in instructional design and technology at Idaho State University. He has a deep fascination with artificial intelligence, particularly its ability to provide effective, supportive, adaptive, differentiated, theory-based learning for all students, regardless of location.

Objectives

Purpose of the Study

The purpose of this study is to develop and evaluate a theory-guided GPT-based Math Tutor carefully and studiously crafted to support community college students preparing to take math placement tests. This project falls under Richey and Klein (2007)'s Type 1 Design and Development Research (DDR) framework, focusing on the design process and formative expert evaluation phases. The study is intended to culminate in the identification of design principles to further enhance the product's effectiveness.

Research Questions

The following research questions guide this study:

RQ1: To what extent does the current Math Tutor GPT design reflect established principles of evidence-informed Tutoring, as assessed by subject matter experts?

Type 1 design fidelity is instrumental for effective GPT design and is the prerequisite for efficacy claims, and is the formative-evaluation deliverable (Richey & Klein, 2007).

RQ2: What design refinements do expert reviewers recommend to improve the pedagogical fidelity and theoretical alignment of the Math Tutor GPT?

This question converts expert reviewers' judgment into actionable revision guidance. It establishes the connection between evaluation and iteration that defines DDR methodology, which treats research as generative rather than strictly corrective.

Significance of the Study

While existing literature addresses many aspects of LLM GPT design for learning, a meaningful gap appears at the intersection of instructional design, educational

technology, learning theory, and generative AI. The Math Tutor GPT is a theory-guided Tutoring tool that is available to students who wish to reinforce their math skills prior to engaging in high-stakes placement testing at the College of Western Idaho. The tool's design framework contributes replicable principles that can be used across institutions. This allows Richey and Klein (2007)'s DDR framework to be applied to similar AI-mediated tools, which are not well addressed in existing literature.

The Math Tutor GPT is available for no cost to either the College of Western Idaho or the students. Additionally, the development cost was fully borne by the researcher, which means that it is freely available to learners beyond the CWI environment. The product is intended to be available at the lowest possible price point, which means that institutions using a GPT are advised to subscribe to ChatGPT. As we will learn from the research results, using the Math Tutor with a free ChatGPT account does not always allow the user to sustain a longer conversation due to context window limitations. A ChatGPT Enterprise license is not required to fully engage with the product.

The CWI Math Solutions Center initially requested the development of the Math Tutor GPT product, which was delivered earlier this year. The CWI Math Solutions Center provides Tutoring support for developmental mathematics to CWI students. The Tutor is intended to mitigate pressure on MSC instructors and staff, allowing them to plan more efficiently.

This Type 1 study intends to produce evidence, through expert review, that will inform subsequent product modifications and upgrades. A subsequent learner-facing study will provide efficacy before the final commissioning of the Math Tutor GPT

product. However, learner outcomes are deferred to the researcher's doctoral dissertation and remain outside the scope of this study.

Scope and Delimitations

This study is limited to remedial mathematics aligned with CWI math placement assessments, though does not address all of CWI's placement tests. However, the GPT instruction framework is designed to automatically expand to include new placement tests when they are added. However, the product currently supports the following placement tests:

- Placement Test Review Unit 0: "Basic Operations on Whole Numbers, Integers, Decimals & Fractions, Basic Algebraic Expressions & Equations."
- Placement Test Review Units 1&2: "Linear Equations and Algebra Basics, Geometry."

The Math Tutor's impact on attitude, learning performance, and organizational results is beyond the scope of the current research. This exclusion is explicit and intentional.

The study was conducted with eight expert reviewers, including mathematics faculty and professional instructional designers. The study proposes to validate the Math Tutor GPT rather than test its efficacy with learners. This is the expected framework for Richey and Klein (2007)'s guidelines for Type 1 product research, and is further addressed in the methodology section of this document. This study is limited to a fidelity review of expert reviewers' interactions with the GPT product.

Theoretical Framework

Effective GPT design demands careful consideration of interlocking theoretical frameworks. Without these, learners will be using an eager, helpful, but untrained and mostly ineffective product. Three theoretical pillars ensure that the Math Tutor remains pragmatically effective. Additionally, they must be operational and sufficiently apparent for reviewers to test and comment on. Therefore, the Tutor's framework stands on a tripod composed of Sweller (1988)'s Cognitive Load Theory, Vygotsky (1978)'s Sociocultural Theory, and Bloom (1968)'s Mastery Learning Model.

The Zone of Proximal Development enhances interactions by generating carefully constructed scaffolded feedback that offers information slightly more advanced than the learner's own. Pacing and the overall organizing principle are provided by Cognitive Load Theory, which manages pacing, sequencing, depth of hints, and worked examples. Finally, the Mastery Learning Model generates a criterion-based framework further enhances pacing while controlling domain progression.

The Math Tutor is coded to integrate seven high-level evidence-based frameworks. To interact effectively, the GPT must integrate and operationalize the contributing learning theories, transforming them from theoretical constructs into concrete, operational aspects. The constructs are explained below.

1. **Diagnostic Questioning:** Corbett and Anderson (1995) knowledge tracing, and Shute and Towle (2003) adaptive instruction.

Tutoring effectiveness is vastly improved by enabling the GPT to accurately apprise itself of the learner's current level of knowledge. This information allows the AI to understand prior knowledge and apply the knowledge-tracing conditions in Shute and

Towle (2003). An initial skill assessment is a common-sense precondition for any sequencing that follows. The Math Tutor begins new sessions by first asking the learner to select the desired pretest, then routing the learner to the starting point that suits them rather than engaging in a blind learning sequence. This helps ensure the student's time is used effectively.

2. **Scaffolded Hint Ladder:** Vygotsky (1978) zone of proximal development, Belland (2017) scaffolding fading, and VanLehn (2006) ITS hint architecture.

In his work on the Zone of Proximal Development, Vygotsky (1978) suggests that hints are most effective when they land just beyond what the learner knows or can manage. The GPT partially assumes the role of a knowledgeable peer. The learner is also protected from being hurried into content that they may be unprepared for, which is a practice suggested by VanLehn (2006) in his work on Intelligent Tutoring Systems (ITS). Given responses from the learner, the Tutor will proceed from a conceptual nudge to a procedural hint, and finally to a full worked step. The idea is to enable a productive struggle without allowing the learner to stall.

3. **Worked Examples with Fading:** Sweller and Cooper (1985) for worked examples effect, Renkl and Atkinson (2003) example-fading sequence, and Sweller et al. (1998) Cognitive Load Theory.

Common sense suggests that a learner who is handed a fully worked example learns more rapidly than one who is handed an unfamiliar problem and told to sort it out. This idea traces to Sweller (1988)'s Cognitive Load Theory and Sweller and Cooper (1985)'s worked examples effect, but is informed by Renkl and Atkinson (2003)'s work on example-based learning. The Math Tutor begins by demonstrating a fully worked

problem, then shifts to a partially completed one that requires the learner to solve a step or two. Finally, the learner advances to independent problem solving. This fading is intended to reduce extraneous load.

4. **Mastery Check Gating:** Bloom (1968) mastery learning, Guskey (2007) criterion-referenced progression, and Corbett and Anderson (1995) knowledge tracing and mastery detection.

Bloom (1968) suggests that the indication of competence is observed when a learner correctly applies a skill before attempting to apply it elsewhere. When paired with Corbett and Anderson (1995) criterion-referenced progression, the principle helps to protect the learner from being hurried into unmastered content. The Math Tutor prevents advancement until the learner has demonstrated proficiency with the current concept.

5. **Spaced Retrieval Prompts:** Cepeda et al. (2006) spacing effect and Roediger and Karpicke (2006) test-enhanced learning.

Cepeda et al. (2006) indicates that retrieval practice distributed across time improves transfer and memory. This idea is enhanced by the integration of fuzzy logic, which periodically revisits and tests previously learned concepts, especially areas of weaker understanding, to reinforce learning (Chrysafiadi & Virvou, 2012). Revisitation consists of small retrieval prompts. Content is relearned until it settles.

6. **Error Classification & Remediation:** Hattie and Timperley (2007) feedback theory, Anderson et al. (1995) cognitive Tutor feedback model, with Booth et al. (2013) error analysis in mathematics. (Note: in the original design specification, this slot was Self-Explanation Prompting; the deployed instrument moved Error Diagnosis up to slot 6 and renamed it Error Classification & Remediation. See methodology note in Phase 2.)

When the learner submits an incorrect answer, the Tutor classifies the error as conceptual, procedural, or careless, and matches the response to the error type rather than issuing undifferentiated "incorrect, try again" feedback (Hattie & Timperley, 2007; Anderson et al., 1995). Conceptual errors receive explanatory depth that targets the underlying principle; procedural errors receive step-level corrective guidance; careless errors receive a brief correction. The classification step itself is the design move that distinguishes a Tutor from a generic chatbot.

7. **Motivational Framing:** Dweck (2006) growth mindset, Dweck and Leggett (1988) implicit theories of intelligence, with Hattie and Timperley (2007) feedback theory.

The Tutor frames the learner’s effort and errors with growth-oriented language, normalizes struggle as part of learning, and reframes mistakes as information rather than failure (Dweck, 2006). The motivational stance is encoded as a turn-level rule: praise effort and process rather than fixed ability; explicitly acknowledge that the current concept is hard and that practice closes the gap; and avoid evaluative language that signals permanent inability. The behavior is grounded in Dweck and Leggett’s (1988) implicit theories of intelligence and operates alongside Hattie and Timperley’s (2007) feedback theory by ensuring that even corrective feedback is delivered through a growth-mindset frame.

Type 2 Model: Scaffolded LLM Tutoring Design Framework (SLTDF)

Educational theory teems with frameworks, each purporting to address some issue or enhance other research, and so on. This project demands a new framework, which is called the Scaffolded LLM Tutoring Design Framework (SLTDF), a Type 2 model that

underlies the Math Tutor GPT. It synthesizes ten learning-science constructs, each designed into the GPT instruction set, including:

- Cognitive load theory (Sweller, 1988)
- Mastery learning (Bloom, 1968; Guskey, 2007)
- Direct instruction, deliberate practice, error-based and diagnostic teaching, gradual release, and scaffolded fading (Belland, 2017; Vygotsky, 1978)
- Retrieval and spaced practice (Cepeda et al., 2006; Roediger & Karpicke, 2006)
- Behaviorist skill fluency (Binder, 1996)
- Transfer-oriented design (Perkins & Salomon, 1992)

These GPT-encoded constructs are translated into seven operationalizable Tutoring behaviors described in the theoretical framework section of the document. Together with GPT session-level interaction rules, including pedagogical stance, refusal patterns, jailbreaking protection, and escalation logic, these guidelines govern how the behaviors behave throughout the interaction. Practitioners can now offer learners a highly consistent, controllable didactic experience. This forethought and coding expertise override and enhance the default helpfulness-seeking behavior of foundation LLM models. The Type 1 expert review forms the initial validation of the SLTDF with the Math Tutor representing the framework's first instantiation. The framework for the SLTDF is shown in Figure 1.

(Figure 1. SLTDF synthesizes ten learning science constructs into seven operationalizable LLM tutoring behaviors that are instantiated in the Math Tutor GPT.)

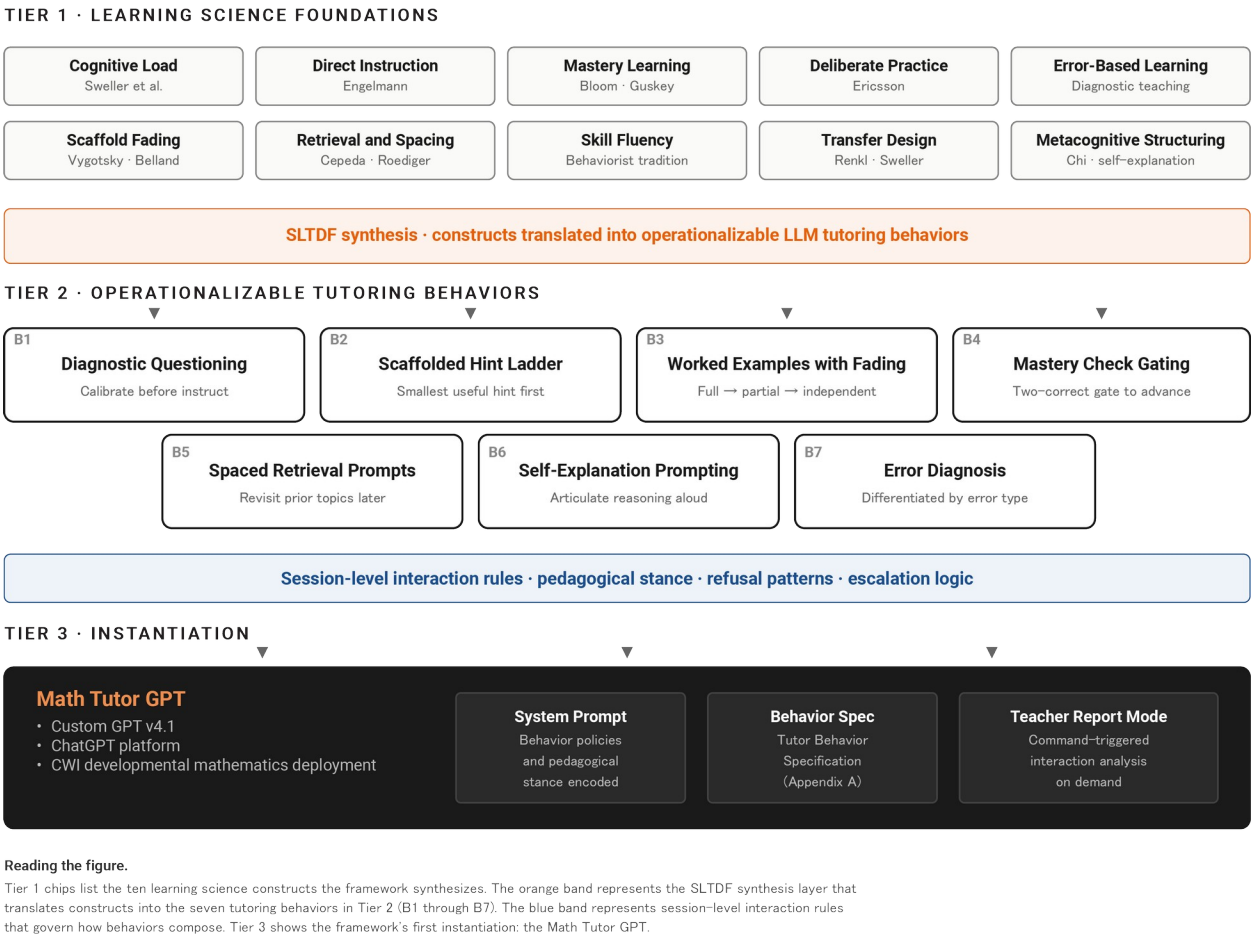


Figure 1. SLTDF synthesizes ten learning science constructs into seven operationalizable LLM tutoring behaviors that are instantiated in the Math Tutor GPT.

Note: Architecture of the Scaffolded LLM Tutoring Design Framework (SLTDF)

Methodology

Research Design Rationale

Richey and Klein (2007) differentiate Type 1, design and development research, from Type 2, model research.

Design Objectives

The Math Tutor GPT operates along five operational dimensions, each linked to a downstream evaluation source. The first consideration is design fidelity, which marks the degree to which the Tutor accurately demonstrates each of the seven evidence-informed Tutoring behaviors (see Appendix D). Expert reviewers judge this aspect against the behavior specification. Secondly, reviewers evaluate conversational coherence to test GPT's pedagogical consistency, a process that requires numerous turns to identify patterns in refusal, hint pacing, and mastery gating, confirming that the model does not revert to generic chatbot behaviors. Third, evaluators are instructed to identify hallucinatory mathematics (or other hallucinatory behavior), spurious procedures and processes, or conversational drift that exceeds or subverts the Tutor's guidelines. Fourth, the Tutor is intended to be a free tool for learners, or as free as possible. In this category, evaluators using free ChatGPT accounts are asked whether the free session offered enough context usability to complete an entire learning session. Finally, testers are asked to address the asynchronous, self-directed, device-independent nature of the product. This Type 1 review directly address the first three dimensions; cost-effectiveness and practical feasibility are addressed, but remain beyond of the scope of this study.

Participants and Settings

Expert reviewers, with disparate though occasionally overlapping expertise in mathematics instruction, professional instructional design, and experience with AI and LLM tool design, participated in the study. All participants are currently employed by the College of Western Idaho. Participants were selected through an emailed invitation based on domain expertise and demonstrated expertise in one of the three domains. Maximum

variation was sought, though only 8 of the 10 invitees participated in the study. All reviewers accepted their assignments voluntarily.

Reviewers participated asynchronously, with no specific testing environment identified. After receiving invitation emails, participants used a hyperlink to navigate to the CWI Math Tutor GPT Design & Development Study webpage. The webpage contains information about the study, about the role of expert reviewers, an informed consent clause, information about the tutoring behaviors and scenarios, instructions, and step-by-step instructions. Hyperlinks to the CWI Math Tutor GPT and the Qualtrics survey pages lead the reviewers to those resources.

Since the researcher is employed at CWI as an instructional designer, the logical decision was made to conduct the study solely at that institution. Several community colleges were considered as potential Type 1 review sites. However, given that CWI requested the creation of the Math Tutor, the researcher did not solicit further institutions. Additionally, CWI has an existing placement testing program. Subsequent research will be carried out with additional institutions that conduct placement testing.

Phase 1: Situation Analysis

Design decisions are necessarily founded upon the needs of the institution, particularly mathematical remediation for struggling learners. Contextual justification is granted to design decisions, which offer both empirical and contextual grounding.

Situational analysis, both empirical and contextual, provides justification for the product design. Design decisions must be focused on the needs of the learner, the literature, and the institutional context, though a novel product may lack academic precedent. The research bridges the gap between validated methods for converting

evidence-informed tutoring principles into controllable and consistent behavior in prompt-engineered GPT-based tutors. In this case, the target population are students who are taking math placement tests.

Phase 2: Design and Development

The Math Tutor specification, or Design Architecture document, is reproduced in Appendix D. This document maps the entire design thought and intent for the development of the Tutor's most recent version. Every evidence-informed tutoring principle is mapped to GPT behaviors that are both auditable and observable. The design fidelity and evaluation criteria demand observable behaviors.

Seven evidence-informed design features are incorporated into the Math Tutor GPT. These include:

1. Diagnostic questioning, which attempts to discern the learner's knowledge at the beginning of a session.
2. Scaffolded hint ladder, meant to provide graduated support from conceptual understanding to procedural confidence.
3. Worked examples with fading, intended to move from fully worked problems to independent practice. Implemented.
4. Mastery check gating, checking for competency before allowing advancement. Implemented.
5. Spaced retrieval prompts, based partially on fuzzy learning, revisit areas of perceived student weakness to reinforce learning (Chrysafiadi & Virvou, 2012). Implemented, but not fully effective, as reported by reviewers.

6. Error classification and remediation, which classifies error types (conceptual, procedural, careless) and provides feedback matched to the type. Implemented.
7. Motivational framing, which uses growth-oriented language, normalizes struggle, and reframes errors as learning opportunities. Implemented.

The researcher is the sole designer, developer, and prompt engineer for the Math Tutor GPT study. All design decisions, including feature creation and modification, prompt engineering modification, and practical theoretical encoding are documented in an ongoing version log. Other prompt engineers can reconstruct the design rationale without needing to contact the researcher.

Additionally, the Math Tutor GPT is designed with a command-triggered teacher report mode, which is activated with the command “show teacher report.” This was not part of the Type 1 expert review. Instead, the review focused on scenario-based evaluation. The teacher mode is intended for future data collection purposes.

Phase 3: Formative Expert Review

This study engages in a structured review process as the primary formative product evaluation method. Type 1 DDR evaluates whether the product displays the design specifications and the theoretical basis of its construction.

Expert reviewers perform evaluations using scenario-based protocols. Each scenario targets one of the seven Tutoring behaviors. Each reviewer holds expertise in at least one expert domain, such as mathematics instruction, instructional design, and prompt engineering. Scenarios include an objective, instructions, and observation criteria. Scenarios do not have to be worked through in a specific order, but it is more efficient to work from first to last. Experts rate their experiences after or during a scenario.

Eight expert reviewers were directed to a protocol webpage that outlined the study's purpose and the theoretical classifications built into the Tutor product. The website included informed consent language, described the seven evidence-informed design features, included step-by-step instructions, and presented the three scripted scenarios. The webpage also includes direct hyperlinks to the Math Tutor GPT and the Qualtrics survey (Appendix F).

Phase 4: Revision

Revision decisions are made following the expert review, which is intended to highlight what works well and what does not. Survey responses shown as “not present” are flagged for critical evaluation. In the case of Math Tutor GPT, reviewers were generally dissatisfied with the spaced-retrieval feature. Revisions stemming from appropriateness ratings below 3.0 will be prioritized for the next iteration of the Tutor. The researcher has carefully maintained versioning logs to preserve design history and traceability.

Phase 5: Design Principle Extraction

Following the expert-informed Phase 4 revision process, principles are extracted based on three pieces of information. First, the structured reviewer ratings and open-ended comments in the survey. Second, prior revisions and logical rationale from the version log. Third, the Tutor behavior specification and seven theoretical foundations from the SLTDF framework. Design principles are grounded in learning science and designed to be generalizable beyond Math Tutor GPT. At least one Type 1 finding is required, such as a rating pattern or a revision decision.

Each recommendation from open-ended comments is sorted by feature, the type of issue that the reviewers encountered, and the framework from the data analysis plan. Issues reported by two or more reviewers constitute a theme. (Richey & Klein, 2007) recommend the following formula: “When designing X, Y should Z because <theoretical rationale>, as evidenced by <evaluation finding>.”

Not all reviewer comments warrant the generation of design principles. For example, reviewer R6 was generally dismissive of the entire process and later reported that they didn’t know how to use ChatGPT, and didn’t complete the entire review. Should design principles be created on the based on the skewed data from this single individual? Though I ran a sensitivity analysis using N-1=7, I’ve decided to process all available data. If N=7 with the outlier excluded, global rating shifts upward by +0.12 to +0.26 per behavior, with overall fidelity shifting from 3.00 to 3.29, the direction of the findings remains unchanged. So, for this study, N=8.

Data Collection Infrastructure

Qualtrics Survey

This Type 1 study collected data exclusively through a Qualtrics survey hosted on Idaho State University’s platform. Reviewers accessed the survey through a hyperlink on the reviewer protocol webpage. Survey modalities include Likert-scale ratings and open-ended responses, all of which were collected by Qualtrics and downloaded as an Excel file once the survey closed. The survey was available to reviewers for approximately one week. Survey data is retained in accordance with ISU data retention policies. Retrospective data is not included in this study.

Math Tutor GPT Session Interaction Transcripts

While it is possible to share ChatGPT sessions, these transcripts were not maintained or requested, and are beyond the scope of this study. Reviewer observations and commentary were exclusively obtained from the Qualtrics survey. A Type 1 DDR study is intended to elicit reviewer judgment about design adequacy rather than an analysis of the Tutor's output. Transcript-level data acquisition will be collected in future learner-facing studies.

Data Analysis Plan

Type 1 DDR research engaged in a convergent mixed-methods design with both qualitative and quantitative elements. A single instrument, a Qualtrics survey, collects both types of data from the same reviewers. This data is used to inform product revision. The survey amasses expert reviewer analysis of the seven evidence-based Tutor behaviors through scenario-based exploration.

Quantitative Analysis

A panel of expert reviewers (n = 8) tested the Math Tutor GPT. Given the small number of quantitative data from the study, it will be descriptively summarized. Four data elements are collected. First, behavioral presence ratings (Yes, Partial, No) in Section B. These include computed frequency counts and percentages for each of the seven design features. Second are behavioral appropriateness ratings (1 = not appropriate to 4 = fully appropriate). This data is analyzed to compute mean, mode, median, standard deviation, and range by feature in Section C. Lastly, global ratings, Section C, and overall design fidelity, Section D, use the same descriptive statistics.

A precent agreement and weighted ordinal statistic, such as Krippendorff's alpha, based on the distribution of responses. Exploratory statistics are preferred over

inferential, considering the N size of the expert panel. This project compiles and generates both qualitative and quantitative elements.

Qualitative Analysis

The directed content analysis is applied to the open-ended elements in Sections B, C, and D. The seven Tutoring behaviors are used as a priori codes taken from the design specification in Appendix D. The scenario-based recommendations for product revision (in Q13, Q17, Q21, Q25, Q29, Q33, Q37), the global comments (Q39), and the three fidelity items for strengths (Q41) and weaknesses (Q42), and the more critical revisions (Q43) are listed by behavior and issue type (presence, appropriateness, implementation, or out-of-scope), and given a priority level based on reviewer language. Inductive codes are added for intersecting themes that recur across behaviors while coded portions are organized in a code book with the dataset.

Analysis Integration

Convergence and divergence of both qualitative and quantitative data, drawn from the seven Tutoring behaviors, is obtained from the scenario-based behavioral ratings. Convergence is identified when a low rating is accompanied by a consistent recommendation for revision. Divergence is identified when specific implementation concerns are identified for the same conceptual item. Research Question 1, design fidelity vs established principles, uses the global ratings in Section C. Research Question 2, expert-recommended product refinements, uses the open-ended form responses in Sections B, C, and D. Expert reviewer profile information is used to stratify the domain of expertise (Section A), but is not considered a primary analytic source. The survey instrument was devised to align item-by-item with the seven Tutoring behaviors and the

two research questions. No further demographic, affective, or learning performance data were collected in this Type 1 study.

Tools and Techniques

This study is comprised of five tools and one artifact supporting the Type 1 expert review. The Math Tutor GPT is deployed as a custom GPT on OpenAI's ChatGPT platform. Reviewers accessed version 4.1 of the Math Tutor GPT using an interface of their choice. Data about the platform, whether a computer, tablet, or mobile device, was not collected. The Tutor works with both unpaid (free) and subscriber-based accounts, although limitations regarding free accounts are noted in the findings.

The study was approved for exempt status by the Idaho State University (ISU) IRB and hosted on ISU's Qualtrics deployment, using that institutions branding and license. Data is retained as-is, and does not contain personal data other than information freely volunteered by the reviewers. The data does not include reviewers' names or other personally-identifiable information, but does contain information about their domain expertise.

Appendix D outlines the Tutor Behavior Specification, which was used as the design document for the development of the GPT product. Numerous prior versions of this document exist, but only the specifications for Tutor version 4.1 are used in this study. The Tutor Behavior Specification is used to compare reviewer experiences to the intended GPT behavior. This is the basis for the analysis of RQ1.

Appendix E, the Expert Review Instrument, contains the Qualtrics survey. The survey collects data for the following elements: structuring presence, appropriateness,

and revision recommendations. The Instrument collects both Likert and open-ended survey information.

Appendix F replicates the reviewer orientation webpage. Reviewers received a link to this website via email invitation. The webpage contains links to the Math Tutor GPT and the Qualtrics survey. The page serves as a comprehensive brief that orients reviewers to the purpose of the project, introduces the seven-behavior brief, three scripted scenarios, the IRB determination, informed consent statement, review process information, and contact information for the reviewer and ISU's IRB.

Appendix B contains formatted information from the Microsoft Excel file used to compile raw results, generate descriptive statistics, and produce charts derived from the Qualtrics CSV export. Spreadsheet tabs include:

- Raw Qualtrics Data: Downloaded as a CSV from the Qualtrics survey.
- Data Dictionary: Converts the raw Qualtrics data into comprehensible, easy-to-read tables.
- Statistics: A quantitative analysis worksheet that includes reviewer demographics, behavioral presence, global ratings, fidelity distribution, score frequencies, and reviewer-by-behavior ratings.
- Charts: Visual display ratings by behavior, presence distributions, fidelity counts, and stacked rating distributions.
- Sensitivity Analysis: Analyzing the possible exclusion of one low-scoring reviewer's responses with incomplete interaction with the Tutor. Tests whether the conclusions remain stable when N = 7 instead of N = 8.
- RQ Alignment: Connects the results to the research questions.

Bias Mitigation

The researcher is both the designer and developer of the Math Tutor GPT. Measures were taken to reduce researcher bias in the expert review findings. First, reviewers were recruited using purposive sampling with expertise in any of three desired expertise domains: Mathematics instruction, instructional design, or experience with AI and GPT development. No single domain dominated the results. Reviewers engaged with the scenarios independently, and to the best of the researcher's knowledge, were unaware of the identities of other reviewers. However, reviewers were afforded the opportunity to be acknowledged by name or remain anonymous in the final report in Appendix E, although none submitted this information.

Type 1 triangulation is accomplished through expertise diversity rather than methodology. All eight reviewers reported at least one domain of expertise, with many identifying multiple domains. Convergence across domains is treated as a primary validity signal. Possible researcher bias limitations are acknowledged in the Limitations section.

Ethical Considerations

Reviewers participated voluntarily and were provided with informed consent in the both the Reviewer webpage and at the beginning of the Qualtrics survey. ISU's IRB determined that this study qualified for exempt status under record IRB-FY2026-207. No personally identifiable information, beyond domain expertise, was collected. Reviewer identities are not associated with the project data.

Reviewers were informed through the reviewer protocol webpage (Appendix F) of their right to withdraw without consequence or explanation, and that submission of the

form constitutes informed consent for participation. All collected data is scrubbed of personally identifiable data, although reviewers could self-identify, if desired. However, aggregated data is reported in de-identified form. Data is retained according to ISU policies.

Discussion and Conclusion

Results and Findings

The survey opened on April 6, 2026, and concluded on April 12, 2026. Ten expert reviewers engaged with the instrument during that period. However, only eight completed the survey before its conclusion. The researcher tested the survey before publication on Qualtrics to ensure full functionality, though that result is not included in the terminal dataset.

Results are organized into four parts. First, reviewers engage with the behavioral presence aspects of the Tutor, identifying consistently observed, inconsistent, and unobserved behaviors. Second, behavioral appropriateness is summarized as a mean, standard deviation, or range, with each feature rated on a 1-4 scale. Percentage values are represent reviewer agreement, given the small N size. Thirdly, relating to Phase 4, Sections B and C afford the opportunity for revision suggestions and are organized thematically. Each feature is organized by feature, type (such as clarity, theoretical alignment, scope, technical), and triaged by revision priority. Fourth, the reliability of the design is characterized, synthesizing the presence and appropriateness of the data with the Section C ratings and cross-observations from Section D. Findings are reported on a feature-by-feature basis. Reviewer profile data from Section A is employed to stratify

findings by expertise domain wherever panel size permits, and is otherwise descriptively reported. Tables and charts related to the findings are reproduced in Appendix A.

Reviewer profile. Ten reviewers engaged with the instrument; eight submitted results (n = 8). Reviewers were invited based on various realms of professional expertise, primarily distributed among mathematics instruction, instructional design, and AI expertise. Many reported more than one area of expertise, which breaks down as follows:

Primary Expertise Area	Count	%
Instructional Design	5	62.5%
Mathematics Education	4	50.0%
Community College Education	4	50.0%
Learning Science	1	12.5%
AI / LLM Tool Design	2	25.0%
Remedial Mathematics	1	12.5%
Student Success	1	12.5%

Reviewers also reported their familiarity with ChatGPT or other LLM models. Seven of eight reported at least some familiarity with LLM interactions, while one reported having no prior experience. The breakdown is as follows:

LLM Familiarity Level	Count	%
No prior experience	1	12.5%
Some familiarity (used, not designed)	3	37.5%
Moderate (designed or evaluated)	2	25.0%
Extensive (research or development)	2	25.0%
Total	**8**	**100.0%**

All reviewers are employed by the College of Western Idaho, with years of experience ranging from 3 to 30. Maximum variation in sampling across mathematics, instructional design, and AI/LLM expertise was achieved as planned.

Behavioral presence. Six of the seven specified behaviors were judged Fully Exhibited or Partially Exhibited by 75% or more of reviewers; one (B5: Spaced Retrieval Prompts) was the exception, with only 12.5% rating it Fully Exhibited and 62.5% rating

it Not Exhibited. Per-behavior frequencies (Yes / Partial / No / No Response) and the percentage rated Fully Exhibited were:

SECTION 2: BEHAVIORAL PRESENCE, SCENARIO EVALUATIONS (n = 8 per behavior)						
Behavior	**Yes, Fully Exhibited**	**Partial**	**No, Not Exhibited**	**No Response**	**% Fully Exhibited**	**% At Least Partial**
B1: Diagnostic Questioning	4	4	0	0	50.0%	100.0%
B2: Scaffolded Hint Ladder	8	0	0	0	100.0%	100.0%
B3: Worked Examples w/ Fading	6	1	1	0	75.0%	87.5%
B4: Mastery Check Gating	5	2	0	1	62.5%	87.5%
B5: Spaced Retrieval Prompts	1	2	4	1	12.5%	37.5%
B6: Error Classification & Remediation	5	2	0	1	62.5%	87.5%
B7: Motivational Framing	7	0	0	1	87.5%	87.5%

The Scaffolded Hint Ladder (B2) was the most reliably present behavior in the deployed Tutor. Spaced Retrieval (B5) is the critical fidelity gap.

Behavioral appropriateness. On the four-point appropriateness scale, the grand mean across all seven behaviors was 3.02 (interpreted as falling between 'Mostly appropriate' and 'Fully appropriate' on the instrument's anchors). Per-behavior descriptive statistics, reported as Mean (SD); Median; Range, were:

SECTION 3: GLOBAL BEHAVIORAL RATINGS, DESCRIPTIVE STATISTICS (1–4 scale)								
Note: Numeric values extracted from Likert text (e.g., "3 — Mostly appropriate" → 3). Blank/missing responses excluded from calculations.								
Behavior	**n**	**Mean**	**SD**	**Median**	**Min**	**Max**	**Range**	**IQR**
B1: Diagnostic Questioning	8	3.13	0.64	3.0	2	4	2	0.25
B2: Scaffolded Hint Ladder	8	2.88	0.83	3.0	2	4	2	1.25
B3: Worked Examples w/ Fading	8	3.25	0.89	3.5	2	4	2	1.25
B4: Mastery Check Gating	8	3.25	1.04	4.0	2	4	2	2.00
B5: Spaced Retrieval Prompts	7	2.29	0.95	2.0	1	4	3	0.50
B6: Error Classification & Remediation	8	2.88	1.13	3.0	1	4	3	2.00

B7: Motivational Framing	8	3.50	0.76	4.0	2	4	2	1.00

The strongest behavior was B7 Motivational Framing (mean 3.50) and the weakest was B5 Spaced Retrieval Prompts (mean 2.29, the only behavior below 2.50). B2 Scaffolded Hint Ladder and B6 Error Classification & Remediation tied at 2.88, indicating that behaviors widely judged present (especially B2 at 100% presence) were not always rated appropriate at the same level.

Reviewer-level variation. Seven global appropriateness items are used to generate the mean ratings of reviewers, which range from 1.83 to 3.71 (n = 8). A small spread, based on maximum-variation sampling design, is a feather rather than a failure of agreement. Spaced Retrieval (B5) and Error Classification (B6) are considered to be the weakest behaviors. This reflects the reviewer's consensus on what works and disagreement with what does not.

Open-ended revision recommendations. Open-ended items in Section B and Section D generated eight convergent themes. They are ranked here by reviewer count, based on the number of reviewers who raised the theme.

1. **Hint Scaffolding (B2): Mean = 2.88.**

   5 of 8 reviewers reported that hint scaffolding gives away answers. Hints and prompts often displayed the answer. Learners are intended to engage in a productive struggle while working through the problem, but the answers sometimes appeared before this could take place.

2. **Spaced Retrieval (B5): Mean = 2.88.**

Generally cited as being absent or unreliable. Retrieval prompts did not appear, or failed to appear regularly. Interestingly, some noticed these prompts only after hitting the ChatGPT free-tier message limits.

3. **Error Class (B6): Mean = 2.29.**

   Reviewers generally agreed that the Tutor failed to differentiate between error types. However, procedural and conceptual errors were treated the same.

4. **Mastery Check (B1, B4): Mean = 3.25.**

   Quick strategy checks were sometimes not interactive. The Tutor posed questions, but failed to wait for an appropriate response.

5. **Validation of Incorrect Answers (Reliability)**

   The Tutor occasionally displayed an incorrect answer, leading to potential confusion for learners. Reviewers found that the tutor would often reverse itself, showing a correct answer after displaying an incorrect answer.

6. **Spaced Retrieval (B5): Mean = 2.29.**

   Hints and equations used jargon too advanced for remedial learners.

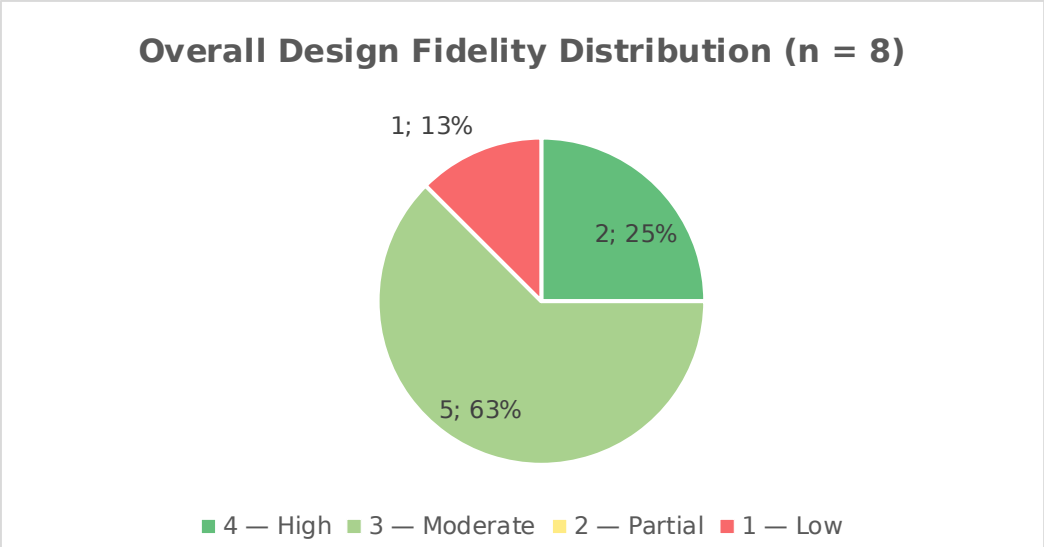
7. **Insufficient Escalation on Repeated Errors (B2, B6): Both means = 2.88.**

   One reviewer reported answering 15+ items incorrectly without noticing escalated assistance from the Tutor.

8. **ChatGPT Free-Tier Limitations (Structural)**

   Reviewers using free ChatGPT accounts reported hitting the free-tier chat limitation wall, which caused premature session truncation.

Overall design fidelity. Aggregate fidelity ratings use a four-point scale (1=Low, 2=Partial, 3=Moderate, 4=High), as illustrated in the following chart:



Among reviewers, 2 (25%) gave the Tutor a high fidelity rating, 5 reviewers (62%) rated it at a moderate fidelity, and 1 (12.5%) rated it at a low fidelity. No reviewers gave the Tutor a Partial fidelity rating. 87.5% of reviewers rated the Tutor at Moderate or High fidelity. The single outlier belongs to the reviewer discussed below.

Sensitivity to outlier exclusion. A single reviewer (Row 8 in the Qualtrics export) reported no prior experience with ChatGPT or LLM-based conversational AI products. The reviewer completed only two of seven scenario-level appropriateness ratings. Furthermore, the reviewer indicated in open response items that they had difficulty navigating the ChatGPT environment. In essence, the reviewer would have to simultaneously learn to use ChatGPT at a fundamental level while analyzing the Tutor. I ran a sensitivity analysis to determine if the N value could be reduced by 1 point, resulting in N = 7 instead of N = 8. After making appropriate calculations to exclude this participant, the global rating means moved upward from +0.12 to +0.26 per behavior, although Spaced Retrieval remained static, since the reviewer left the item blank. Additionally, the fidelity mean rises from 3.00 (at n = 8) to 3.20 (at n = 7), though the

median remains unchanged. Neither n value, 7 or 8, changed the direction of findings. Given the modest n, it was decided to keep the reviewer's responses, even if incomplete.

Synthesis (RQ1 and RQ2). Research Question 1: Extent of Design Fidelity

An overall mean of 3.00 across all reviewers, or 3.29 excluding the outlier, and a grand mean of 3.02 indicate that the Tutor achieves a moderate to high design fidelity. With a mean of 3.50, Motivational Framing (B7) stands out as the strongest realization of the seven desired behaviors, while Spaced Retrieval (B5) is the weakest, with a mean of just 2.29, or largely absent. Six of the seven behaviors are reliably present in the Tutor. The Tutor shows significant evidence of the seven behaviors, although it does not achieve full fidelity to the design specification.

Research Question 2: Recommended Refinements

Triage of open-ended responses led to suggestions for product refinements. Primarily, the Hint Scaffolding, the weakest feature, requires reconsideration and revision. While features are interconnected in various ways, it is most efficient to address one at a time. For the Tutor, modification triage will proceed along three axes: First, reengineering most critical failure points, and second, the most complex and interconnected, and third, the easiest.

Spaced Retrieval is moderately complex, and, given its low mean, will be address first, followed by either Error Classification, which should be easier, and Spaced Retrieval, which is more complex. Along with non-interactive strategy checks, these four will constitute the first targets for modification and testing.

Interpretation

Findings are intrinsically tied to the design rationale established in the theoretical framework. The analysis highlights which of the seven Tutoring behaviors demonstrated the highest fidelity in the reviewers' eyes. Why did the prompt engineering techniques work so well for the highest-rated behaviors? Each relies on explicit conversational rules rather than contextual judgment, and the gap between coding and realization is most evident between principle and instantiation.

Session-level state tracking presents a real challenge in a GPT environment. This is the process that powers the Spaced Retrieval, the lowest-rated method. Rectifying the underperformance of Error Classification will be less challenging, as the modifications are more straightforward and can be directly encoded within the GPT’s core logic. The low scores were surprising to the researcher. It can certainly operate more effectively than it did.

The GPT itself works through modularity, with logical integrations and a sound decision-making core. Interconnections between modules are both common and necessary, and represent cutting-edge prompt engineering. Type 1 research offers a superior avenue for building, testing, and deployment, and does not intend to accrue student experiences. Research about intelligent tutoring systems (ITS) (Anderson et al., 1995) offer value, but it must be acknowledged that most such research was embarked upon before the appearance of Large Language Models, and, although clever in their implementation, were still computer programs, that, although built with significant theoretical foresight, present significant limitations that do not apply to artificially intelligent environments. Each feature presents a unique challenge, yet still has much in common, using the same fundamental prompting language.

Type 1 research focuses on the adequacy of the design, not learner efficacy. Claims about the Tutor's effect on students are beyond the scope of the current study, but should generate an interesting and exciting wealth of data.

Design Principles Generated

Richey and Klein (2007) offer a structured format for stating design principles generated by a DDR project based on this format: "When designing [X], [Y] should [Z] because [theoretical rationale], as evidenced by [evaluation finding]." Nearly mathematical in its potential precision, it works well with the findings from this research project. Five fundamental principles became apparent post-research.

Principle 1: When designing scaffolded hint ladders in an LLM tutor, the system should withhold answer-revealing structure until the learner has attempted the problem, because Vygotskian scaffolding requires productive struggle within the zone of proximal development, as evidenced by five of eight reviewers independently flagging that the tutor's hints sometimes handed the learner the answer rather than guiding to it (hint-ladder appropriateness M = 2.88).

Principle 2. When designing for spaced retrieval in an LLM tutor, the system should engineer explicit session state rather than rely on the model to remember prior content, because the spacing effect requires deliberate reintroduction of material over time, as evidenced by spaced-retrieval being rated fully exhibited by only 12.5% of reviewers and absent by 62.5% (M = 2.29, the lowest of the seven behaviors).

Principle 3. When designing tutor feedback in an LLM tutor, the system should classify error type as conceptual, procedural, or careless before responding and route to type-specific remediation, because process-level feedback yields larger learning gains

than uniform correction, as evidenced by three of eight reviewers flagging undifferentiated error feedback as a refinement priority (error-classification appropriateness M = 2.88).

Principle 4. When designing self-explanation prompts in an LLM tutor, the system should require learner input before proceeding, because the self-explanation effect depends on active learner generation rather than rhetorical pause, as evidenced by three of eight reviewers reporting that strategy-check prompts continued without waiting for the learner's response.

Principle 5. When developing prompt-engineered LLM tutors, designers should produce a written behavior specification before prompt engineering begins, because Type 1 fidelity assessment requires a design contract that maps each principle to an auditable behavior, as evidenced by an overall fidelity rating of M = 3.00 (moderate but not full) and by the researcher's reflection that several reviewer-flagged issues traced to places where the specification was implicit rather than written.

Theoretical Implications

Type 1 research is concerned with the design of the product rather than the downstream effect on learners. The idea is to find ways to enmesh learning principles directly into the AI product, which often requires guile on the part of the designer.

Behaviors that can be engineered into the GPT as turn-based are far easier to encode than those that depend on a session-level actualization. Turn-based behaviors include mastery gating, hint laddering, and refusal patterns. Session-level prompting includes spaced retrieval, diagnostic questioning, and scaffolding. However, the session

state must be remembered by the AI, and based on how a large language model actually works, every feature lives partially in both session and turn-based arenas.

The Tutor is designed to maintain a metacognitive profile of the learner while retaining their current state of ability, their prior knowledge, the immediate next steps, and the end goal. The ability for the Tutor to accomplish this relies heavily on the context window. Context window refers to the maximum amount of text, measured in the number of tokens, that the AI model can process and remember during any point in the conversation, and it is not static. If a model has a context window of 100,000 tokens, then that number is counted backward from the current point of the conversation. Therefore, when conversations exceed that amount, it will begin to forget what happened at the beginning of the session unless specifically prompted to remember. The Tutor includes a mitigation strategy to partially sidestep the effects of a moving context window. Due to the nature of an LLM, this is never perfect, and participants using free ChatGPT accounts noticed the reduced context as they ran afoul of various account restrictions.

Another issue is that the a GPT, becomes less reliable when it must reason over an evolving declarative model, which the Tutor uses in conjunction with procedural functions. Additionally, the researcher desired to apply the principle of a more knowledgeable peer (Vygotsky, 1978), but a GPT's base prompt, designed by OpenAI, is designed to please. This leads to the possibility of subversion despite programming when the user continues to probe the model to achieve something beyond of the GPT coding. This is commonly referred to as jailbreaking. The Tutor has robust jailbreaking protections, but even so, a persistent or clever user, or one who displays emotions that

ChatGPT is natively coded to respond to in a certain way, may still be able to derail the interaction.

Intelligent tutoring systems might provide a solution, but as a concept, they are more concerned with static systems, branching logic, and direct support based on known errors that students commit. In other words, they are not AI systems. The Tutor still draws from such ideas, but it is artificially aware, and the transformer architecture, even at medium temperature and Top P settings, will always exhibit more apparent chaos than an ITS interaction. A capable prompt engineer will desire to compensate for this through prompting. Unfortunately, unless using the API, ChatGPT does not allow direct changes to either Temperature or Top P.

However, large language models are exceptionally adept at understanding and delivering procedural, cognitive, and constructivist learning. But it is best to understand how they work at an intuitive level, which is the product of a great deal of model usage by the prompt engineer. Even the researcher, currently in the top one percent of all users, sometimes struggles to achieve very specific and repeatable interactions and outcomes tied to operating multiple theoretical constructs, often cooperating, but sometimes competing for attention or primacy within the same construct.

Model Implications

While coding the base Tutor prompt to accept and effectively interact between seven different theoretical constructs, it became obvious to the researcher that an overarching framework is required to manage those interactions

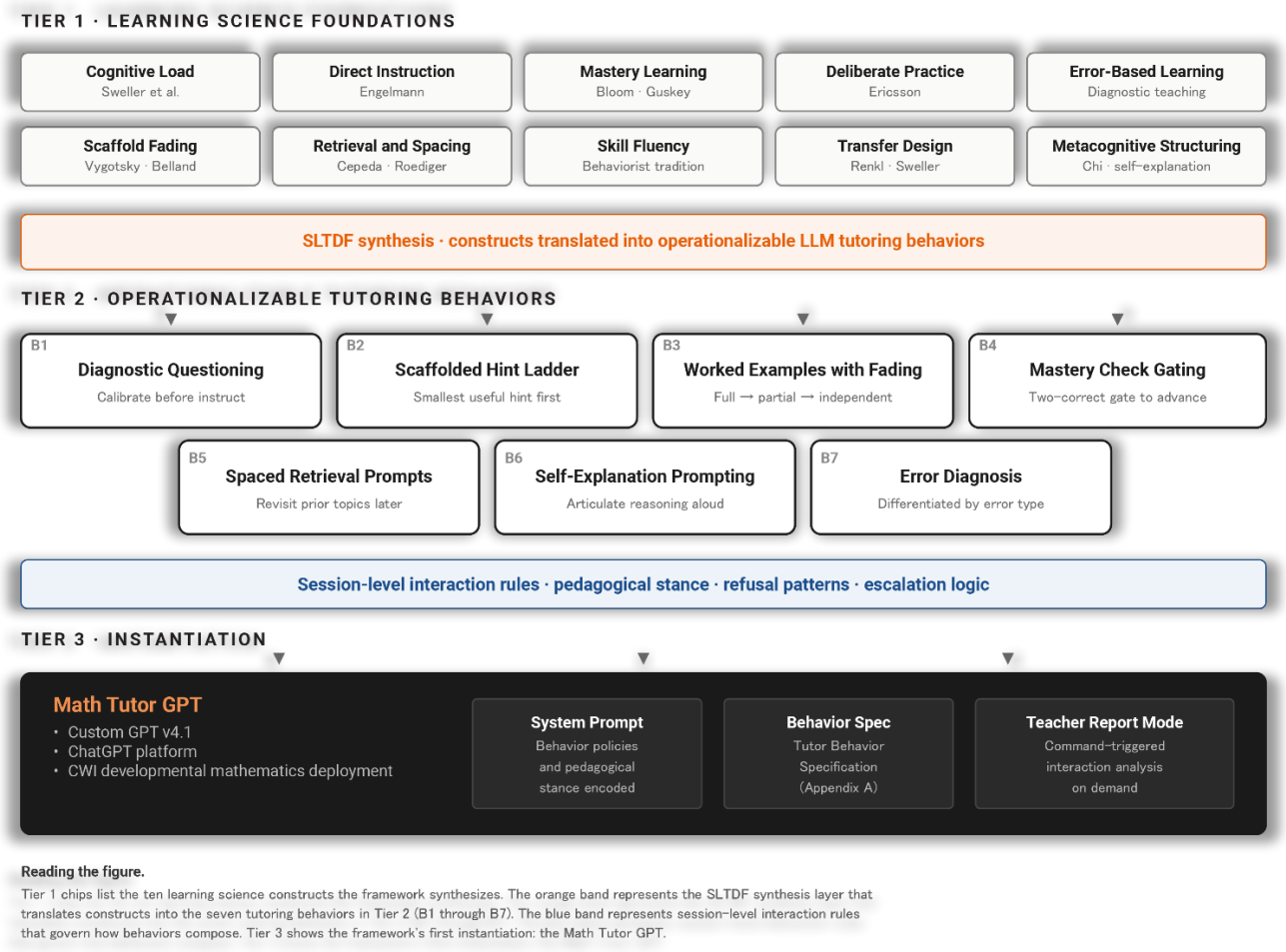
The researcher quickly noticed during the initial development phases of the Tutor that a framework, or guiding model, would be required to effectively manage interactions

among the seven behaviors. This manager is intended to decide which behaviors would be evidenced, and which would be reduced, at any given time in a chat session based on the user's actions and the session state.

This session manager is called the Scaffolded LLM Tutoring Design Framework (SLTDF), and its first instantiation is in the Math Tutor GPT. However, the SLTDF is merely the most recent of an evolving series of such managers, although this is the first deployment in a mathematics-specific tutor (Figure 1. SLTDF synthesizes ten learning science constructs into seven operationalizable LLM tutoring behaviors that are instantiated in the Math Tutor GPT.).

Figure 1. SLTDF synthesizes ten learning science constructs into seven operationalizable LLM tutoring behaviors that are instantiated in the Math Tutor GPT.

SLTDF Model Framework for GPTs



Note: Three-tier diagram showing how ten learning science constructs (cognitive load theory, direct instruction, mastery learning, deliberate practice, error-based learning, scaffold fading, retrieval and spacing, skill fluency, transfer-oriented design, and metacognitive structuring) are synthesized through the SLTDF into seven operationalizable LLM tutoring behaviors (diagnostic questioning, scaffolded hint ladder, worked examples with fading, mastery check gating, spaced retrieval prompts, self-

explanation prompting, and error diagnosis with differential response) and instantiated in the Math Tutor GPT (custom GPT v4.1).

The SLTDF acts as the interpretive layer between the actual learning theories and their practical realization in the model. For example, we know that worked examples with fading is effective, but the system needs to understand, in very straightforward terms, how that is meant to work. So, for example, the SLTDF must to draw from Mastery Learning (Bloom, 1968) and Scaffolded Fading (Vygotsky, 1978), as well as others, to achieve a specific type of interaction.

With this connection between theories and delivery established, the information is passed to the actual Tutoring engine, which combines session-based rules and guidelines, the current pedagogical stance, and the interaction from SLTDF into a pragmatic protocol that can be delivered.

Given the responses from Sections C and D, it becomes clear that the framework's session-level guidance needs to be improved, while the turn-based elements were more robust. It is important, for future Tutor iterations, that the weaker areas be carefully identified, redesigned or modified, and tested extensively. However, the framework was indeed instantiated and clearly recognized by reviewers in the custom GPT. Perfection in GPT design is a matter of balance and judiciousness. The researcher may choose to test multiple frontier AI platforms for future testing, and if the SLTDF model is inseparable from the Tutor, future research might demand a full exposition of Type 1 and Type 2 elements.

Limitations

Seven limitations stand out for this research project:

First: One of the seven Tutoring behaviors originally designed into the Tutor was replaced with Motivational Framing, which the researcher considered an improvement to the product, and is not covered in this study. However, given that this is a design and development research project, interest is focused on whether the behaviors are evidenced to reviewers rather on its effect on learners. This is not a shortcoming, but a scope decision.

Second: With a small n = 8, the study cannot generalize across a larger population. This means that the results are delivered in percentages rather than coefficients that might assume larger samples. Further research is advised to obtain a much larger n size, but, given the time necessities, this luxury was beyond the grasp of the researcher's resources.

Third: The dual or multiple domain expertise roles of reviewers are mentioned, but not woven into the fabric of the actual results, although a more advanced study would benefit from this and other demographic examinations. Curiosity in further research should ask these questions, seeking to determine how demographics influence perception of and interaction with the Tutor.

Fourth: Time constraints presented a serious issue. This project had to be completed within a handful of weeks, and included IRB submission and acceptance, building a reviewer website, creating a Qualtrics survey, creating scenario-based testing, obtaining reviewers, collecting results, analyzing data, creating a presentation, and writing a report. All acceptable practices, but compressed well beyond what a more robust study would demand.

Fifth: The researcher chose to use ChatGPT for the GPT functionality. The other option, a Gemini Gem, may have presented a better option, but the Tutor's design had already begun in ChatGPT in 2025. Still, ChatGPT is a well-known, highly capable and stable platform. The researcher considered moving the research to Claude, an even more capable platform, but Claude doesn't have an equivalent of GPTs or Gems, meaning that the researcher would have to use the pay-as-you-go structure of Claude's API.

Sixth: Intellectual property considerations also play into the decision to continue with ChatGPT. GPT instruction sets (prompts) are concealed from users, and cannot easily be hacked, as they could in the past. The researcher uses numerous routines that, if shared, might require an NDA. This may change, but sharing these routines, which are used elsewhere, would give others the opportunity to publish and claim credit for the researcher's inventions, which is not optimal for the researcher's advancement. However, institutional replication of the protocol is a major consideration that may entirely modify this consideration.

Seventh: Given the tight timing of the project, the researcher was unable to pilot the GPT or survey instruments with a separate set of reviewers beforehand. The debrief in Section D indirectly serves some of these functions, but does not substitute for a formal pilot.

Implications

Concerning generalizability, implications can be described in three tiers. First, for instructional designers at community colleges, who place a high value on transferable process knowledge as an area of contribution. Here, the practice of using iterative prompt engineering is key. Writing specifications before creating the GPT remains important,

offering a preconceived plan to fulfill in the design phases. For example, behavior specification is best considered before coding begins, and their question will be, “what do we wish to accomplish, how should it behave, what theoretical approaches would be most effective in subject [X], and how does everything connect and play out in reality?” All of these processes are generalizable.

Second, and perhaps most importantly, the considerations for AI developers are sharper. Creation of the actual product must be created from many disparate elements. Each of the tutoring behaviors, for example, must possess session-level capabilities, interacting with the other behaviors through the SLTDF interconnect architecture. Here, the weights and precedence are determined and managed. While SLTDF allows the architecture to be reproduced, it also requires fine-tuning. Additionally, the chatbot must be instructed to refuse, withhold, and gate rather than agree, actions that partially subvert the base model prompt. Jailbreaking protection also presents a significant challenge, and effective protection relies on a deeper understanding of the way that an LLM actually works.

Finally, while the actual deployment to students is beyond the scope of this study, further research demands that the actual rollout be both controlled and visible. A secret teacher report mode is designed into the Tutor, accessible only by typing in a specific phrase. This will help to supply actual interaction data. However, the mode was not part of the testing, and none of the reviewers were apprised of its existence. Such outcomes are not part of a Type 1 study.

Lessons Learned

The Math Tutor GPT has progressed through multiple versions, starting before the well research began by the request of the Math Solutions Center, but with the intention of using it for research. The Tutor has moved through the following versions: V0.5, V2, V3, V3.5, V4, and V4.1. Each version's change long has been preserved. However, although the prompt is an engineered product, it

was treated as a singular unit, which was a decision that enabled me to champion specific design choices with expert reviewers.

The Reviewer protocol webpage (Appendix F) was a methodological decision. It framed the entire project for reviewers, outlined the Tutor's features, and defined the testing sequences and scenarios. The choice to test asynchronously appears to have been well-founded, with little to no feedback or questions received while the reviewers interacted with the product on their own schedules. An attractive, organized research introduction has proven valuable. The reviewers know what they are meant to do, learn a little about the theories and features, and proceed without issue.

The decision to focus on scenario-based protocols to inspect specific behaviors has proven to be a good decision. Rather than engage with the tool in an open-ended way, the targeting provided focused interaction.

The largest challenge proved to be limited time. Given adequate time and a more comprehensive plan, the research could progress far more smoothly. Additionally, additional time to pretest with selected reviewers would prove beneficial and instructive.

Recommendations

The first logical next step would be to extend beyond DDR into student trials. Pre- and post-placement tests follow as a natural next step. These seek to compare the

current Math Tutor GPT to an untrained GPT as a baseline product, or a non-theory-infused product, to determine whether the former improves learning over the baseline.

Secondly, an observational study of learners' interaction behavior would reveal how they engage with the Tutor. The value here is undeniable for product research and development. Analogously, it is identical to the processes of video game testing, which is well known in software development.

Third, careful analyses would determine which types of learners benefit most from the Tutor, based on various design aspects. Naturally, this would include examination of initial mathematical proficiency, math anxiety, prior experience with LLMs, and usage intensity, generating differential data rather than suggesting a single-track treatment effect.

Fourth, like any software, the design benefits from iterative, carefully considered development cycles. As a moving target, SLTDF improves as evidence accumulates.

Fifth, SLTDF can be tested at scale with questions or interactions that drift, which is common in LLM. Thousands of interactions can precisely identify areas of strength and weakness.

Sixth, the framework can be replicated across comparable open-enrollment community colleges using more advanced mathematical domains. This would ideally enhance the Type 2 generalizability of the SLTDF framework, as required for DDR model claims (Richey & Klein, 2007).

Seventh, the product should be moved from a GPT environment to a software platform that uses the API from Claude. The experience would still include the chat features, but add a mathematical writing tool, such as MathType. Such a feature would

make it easier for users to enter their answers without attempting to write fractions on a single line, for example.

Conclusion

This Type 1 design and development study is intended to answer two questions. Research Question 1 asked the extent to which the current Math Tutor GPT (v4.1) design exhibits the principles of evidence-informed Tutoring as reported by expert reviewers and subject matter experts. Collected evidence demonstrated that the reviewers were able to identify the seven behaviors in scenario-based testing. The Tutor's fidelity with the behaviors aligns with the theoretical basis. Exhibited variation from the design document is reported in detail.

Research Question 2 solicited the reviewers for design guidance and suggestions for future versions of the tutor, focusing on theoretical alignment and usability. Open-ended questions allowed reviewers to comment and share their thoughts about functionality and adherence to the seven behaviors, which are recorded in the revision log for correction in subsequent Tutor versions.

The current study contributes to subsequent and ample opportunities for research, providing broader generalizations about the framework and Tutor as per Richey and Klein (2007). Learner outcomes are beyond the scope of this study, though further research will remain incomplete without investigating student outcomes.

The contribution provided by this research demonstrated that the Math Tutor GPT operates as intended, and that its design features are recognizable by experts as being theoretically grounded. This represents the first piece of formative evidence in the efficacy of the framework.

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, *4*(2), 167–207.

Belland, B. R. (2017). *Instructional scaffolding in STEM education: Strategies and efficacy evidence*. Springer. https://doi.org/10.1007/978-3-319-02565-0

Binder, C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst*, *19*(2), 163–197. https://doi.org/10.1007/BF03393163

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, *1*(2), 1–12.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). *Distributed practice in verbal recall tasks: A review and quantitative synthesis* (0033-2909). (Psychological Bulletin, Issue. https://research.ebsco.com/linkprocessor/plink?id=2e4a20a6-4a2d-3040-82b9-e6587978a028

Chrysafiadi, K., & Virvou, M. (2012). Evaluating the integration of fuzzy logic into the student model of a web-based learning environment. *Expert Systems With Applications*, *39*(18), 13127–13134. https://doi.org/10.1016/j.eswa.2012.05.089

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253–278.

Guskey, T. R. (2007). Closing achievement gaps: Revisiting Benjamin S. Bloom's "Learning for Mastery". *Journal of Advanced Academics*, *19*(1), 8–31. https://doi.org/10.4219/jaa-2007-704

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husén & T. N. Postlethwaite (Eds.), *International Encyclopedia of Education* (2nd ed., Vol. 11, pp. 6452–6457). Pergamon Press.

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist*, *38*(1), 15–22. https://doi.org/10.1207/S15326985EP3801_3

Richey, R. C., & Klein, J. D. (2007). *Design and development research: Methods, strategies, and issues* (1 ed.). Routledge. https://doi.org/10.4324/9780203826034

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Shute, V. J., & Towle, B. (2003). Adaptive e-learning. *Educational Psychologist*, *38*(2), 105–114. https://doi.org/10.1207/S15326985EP3802_5

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cognitive Science*, *12*(2), 257–285.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*(1), 59–89. https://doi.org/10.1207/s1532690xci0201_3

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227–265.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Appendices

Appendix A: Charts and Graphs

The following charts and graphs visualize the descriptive statistics derived from the expert review responses. They were generated from the Qualtrics export sheet. Each figure includes a brief caption.

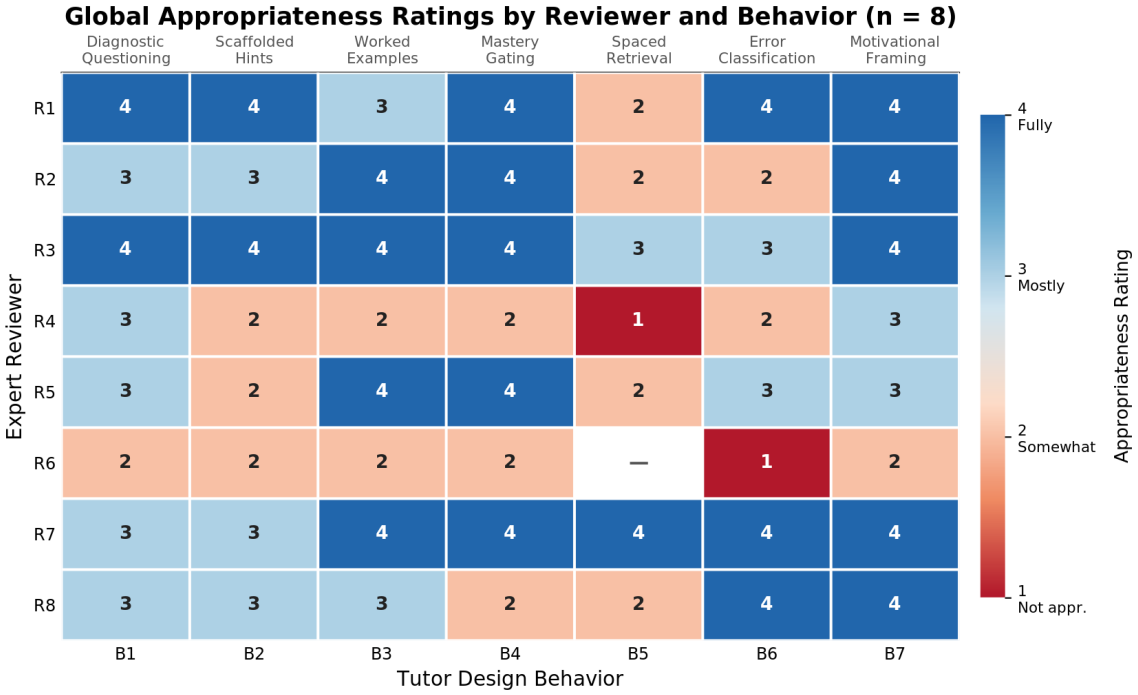


Figure A1. Reviewer Behavior Heatmap.

Note. Heatmap visualization of reviewer ratings across the seven design features and eight reviewers.

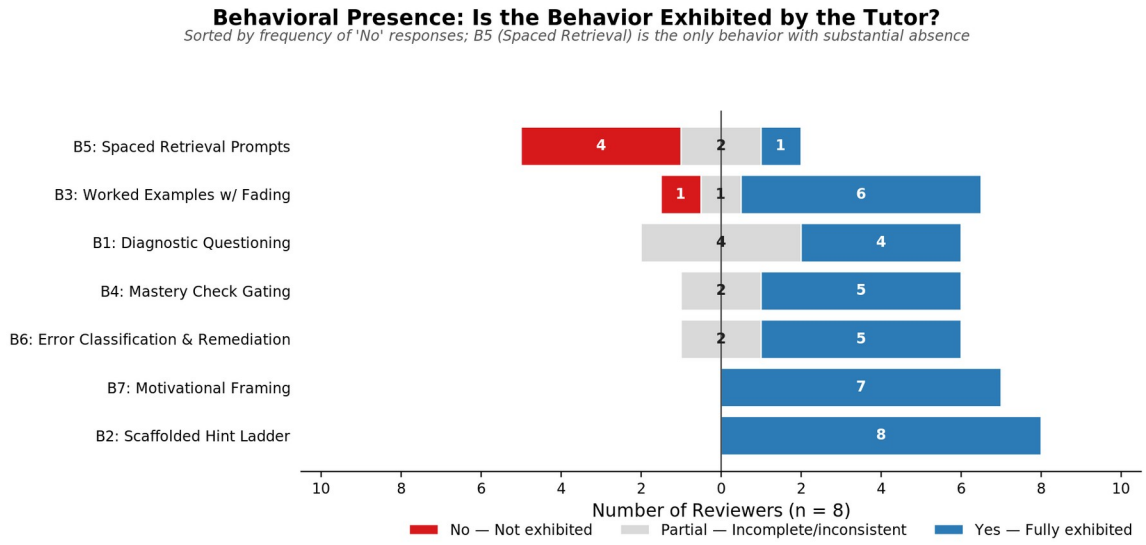


Figure A2. Behavioral Presence (Diverging Bar Chart).

Note. Diverging bar chart showing the proportion of reviewers who judged each behavior as present versus absent.

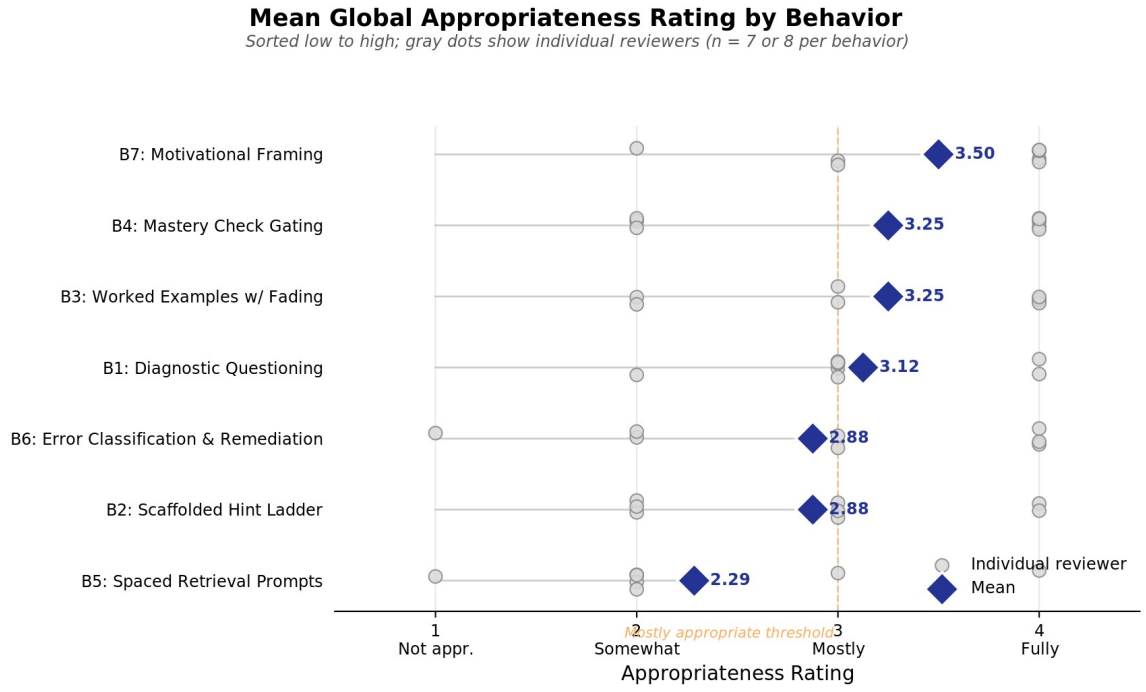


Figure A3. Global Ratings Dot Plot.

Note. Dot plot showing each reviewer's global ratings across evaluation dimensions.

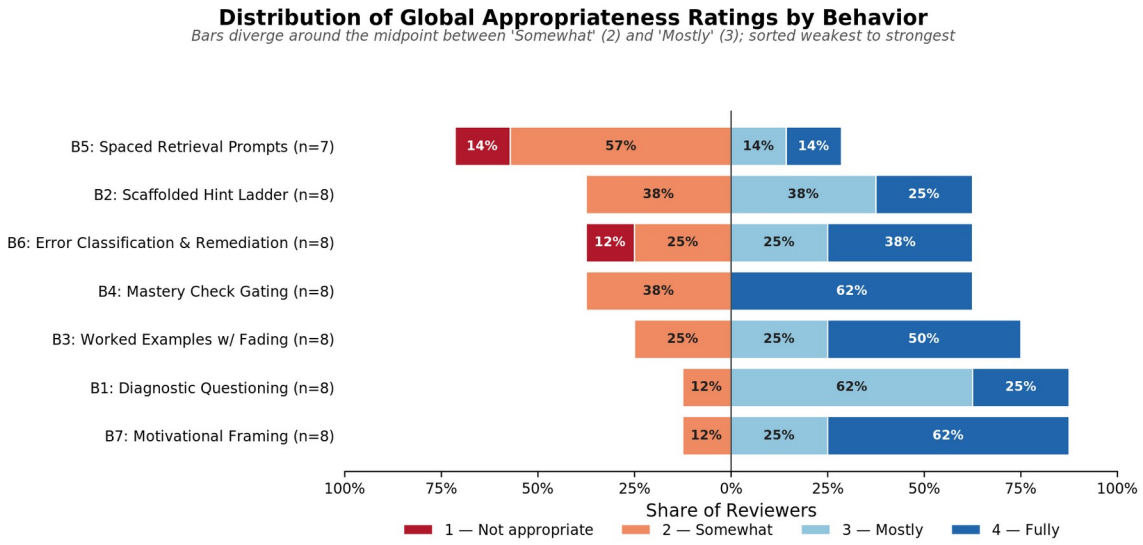


Figure A4. Likert Response Distribution.

Note. Diverging Likert distribution showing the spread of agreement across appropriateness items.

Appendix B: Raw Data

AI Math Tutor Expert Review Instrument - Qualtrics Export

Part 1: Data Dictionary

The following table defines each column in the raw Qualtrics export, including variable name, description, survey section, data type, and response options. This dictionary corresponds to the "Data Dictionary" worksheet in the source Excel workbook. The table continues on the next (landscape) page.

AI Math Tutor Expert Review — Data Dictionary	
SURVEY OVERVIEW	
Instrument	AI Math Tutor Expert Review Instrument (Qualtrics)
Purpose	Expert evaluation of an LLM-based math tutoring GPT against a 7-behavior design specification
Total Responses	8
Total Columns	60
Data Rows in Sheet0	Row 1 = Variable names, Row 2 = Full question text, Rows 3–10 = Responses
Respondent Pool	Faculty, instructional designers, and staff at College of Western Idaho (CWI)
Collection Date Range	April 2026 (Qualtrics export April 9, 2026)

COLUMN-BY-COLUMN DICTIONARY					
Column	Variable Name (Row 1)	Description	Section	Data Type	Response Options / Notes
A	StartDate	Survey start date/time	Qualtrics Metadata	Date/Time (serial)	Excel serial number; format as date to read
B	EndDate	Survey end date/time	Qualtrics Metadata	Date/Time (serial)	Excel serial number; format as date to read
C	Status	Response type (e.g., IP Address, preview)	Qualtrics Metadata	Text	"IP Address" = live response
D	IPAddress	Respondent IP address	Qualtrics Metadata	Text	Masked with ******* in this export
E	Progress	Survey completion progress (%)	Qualtrics Metadata	Numeric (0–100)	100 = fully completed
F	Duration (in seconds)	Time spent on survey in seconds	Qualtrics Metadata	Numeric	Range in data: ~331 to ~13,831 seconds
G	Finished	Whether respondent finished the survey	Qualtrics Metadata	Boolean	True / False
H	RecordedDate	Date/time response was recorded	Qualtrics Metadata	Date/ Time (serial)	Excel serial number
I	ResponseId	Unique Qualtrics response	Qualtrics	Text	Format:

		identifier	Metadata		R_xxxxxxxxxxxxxxx
J	RecipientLastName	Recipient last name	Qualtrics Metadata	Text	Masked with ******* in this export
K	RecipientFirstName	Recipient first name	Qualtrics Metadata	Text	Masked with ******* in this export
L	RecipientEmail	Recipient email address	Qualtrics Metadata	Text	Masked with ******* in this export
M	ExternalReference	External data reference	Qualtrics Metadata	Text	Masked with ******* in this export
N	LocationLatitude	Respondent latitude	Qualtrics Metadata	Numeric	Blank for all responses in this export
O	LocationLongitude	Respondent longitude	Qualtrics Metadata	Numeric	Blank for all responses in this export
P	DistributionChannel	How the survey was distributed	Qualtrics Metadata	Text	All responses = "anonymous"
Q	UserLanguage	Language of the survey	Qualtrics Metadata	Text	All responses = "EN"
R	Q_RecaptchaScore	reCAPTCHA bot-detection score	Qualtrics Metadata	Numeric (0–1)	1 = very likely human; all responses = 1
SECTION A: CONSENT & REVIEWER DEMOGRAPHICS					

S	Q45	Informed consent agreement	A: Consent	Text	"I Agree — I have read the informed consent information and agree to participate"
T	Q3	Professional role / title (optional)	A: Demographics	Open text	Examples: Math Teaching Faculty, Instructional Designer, Higher Education Staff
U	Q4	Institution / organization	A: Demographics	Open text	All respondents: College of Western Idaho (CWI)
V	Q5	Primary area(s) of expertise (select all that apply)	A: Demographics	Multi-select text	Options include: Instructional Design, Learning Science, Mathematics Education, Remedial Mathematics, Community College Education, AI/LLM Tool Design, Student Success
W	Q6	Years of relevant professional experience	A: Demographics	Numeric / Text	Range in data: 3 to 30 years
X	Q7	Familiarity with LLM-based educational tools	A: Demographics	Ordinal text	No prior experience / Some familiarity (have used, but not designed) / Moderate (have

					designed or evaluated) / Extensive (research or development background)
Y	Q8	Conflict of interest disclosure (optional)	A: Demographics	Open text	Disclosure of any relationship to the researcher or CWI that could bias evaluation
SECTION B: SCENARIO-BASED BEHAVIOR EVALUATIONS (S1–S7)					
Z	Q11	S1 — Behavioral Presence: Diagnostic Questioning	B: S1 Diagnostic Questioning	Ordinal text	Yes — fully exhibited / Partial — exhibited but incomplete or inconsistent / No — not exhibited
AA	Q12	S1 — Behavioral Appropriateness: Diagnostic Questioning	B: S1 Diagnostic Questioning	Ordinal (1–4)	1 — Not appropriate / 2 — Somewhat appropriate / 3 — Mostly appropriate / 4 — Fully appropriate / NA (skip if Presence = No)
AB	Q13	S1 — Revision Recommendations: Diagnostic Questioning	B: S1 Diagnostic Questioning	Open text	What should change, and why? (Optional)
AC	Q15	S2 — Behavioral Presence:	B: S2 Scaffolded	Ordinal text	Yes — fully exhibited /

		Scaffolded Hint Ladder	Hint Ladder		Partial / No — not exhibited
AD	Q16	S2 — Behavioral Appropriateness: Scaffolded Hint Ladder	B: S2 Scaffolded Hint Ladder	Ordinal (1–4)	1–4 scale (same as above) or NA
AE	Q17	S2 — Revision Recommendations: Scaffolded Hint Ladder	B: S2 Scaffolded Hint Ladder	Open text	Optional
AF	Q19	S3 — Behavioral Presence: Worked Examples with Fading	B: S3 Worked Examples with Fading	Ordinal text	Yes — fully exhibited / Partial / No — not exhibited
AG	Q20	S3 — Behavioral Appropriateness: Worked Examples with Fading	B: S3 Worked Examples with Fading	Ordinal (1–4)	1–4 scale or NA
AH	Q21	S3 — Revision Recommendations: Worked Examples with Fading	B: S3 Worked Examples with Fading	Open text	Optional
AI	Q23	S4 — Behavioral Presence: Mastery Check Gating	B: S4 Mastery Check Gating	Ordinal text	Yes — fully exhibited / Partial / No — not exhibited
AJ	Q24	S4 — Behavioral Appropriateness: Mastery Check Gating	B: S4 Mastery Check Gating	Ordinal (1–4)	1–4 scale or NA

AK	Q25	S4 — Revision Recommendations: Mastery Check Gating	B: S4 Mastery Check Gating	Open text	Optional
AL	Q27	S5 — Behavioral Presence: Spaced Retrieval Prompts	B: S5 Spaced Retrieval Prompts	Ordinal text	Yes — fully exhibited / Partial / No — not exhibited
AM	QID28	S5 — Behavioral Appropriateness: Spaced Retrieval Prompts	B: S5 Spaced Retrieval Prompts	Ordinal (1–4)	1–4 scale or NA
AN	Q29	S5 — Revision Recommendations: Spaced Retrieval Prompts	B: S5 Spaced Retrieval Prompts	Open text	Optional
AO	Q31	S6 — Behavioral Presence: Error Classification & Remediation	B: S6 Error Classification & Remediation	Ordinal text	Yes — fully exhibited / Partial / No — not exhibited
AP	Q32	S6 — Behavioral Appropriateness: Error Classification & Remediation	B: S6 Error Classification & Remediation	Ordinal (1–4)	1–4 scale or NA
AQ	Q33	S6 — Revision Recommendations: Error Classification & Remediation	B: S6 Error Classification & Remediation	Open text	Optional
AR	Q35	S7 — Behavioral Presence:	B: S7	Ordinal text	Yes — fully exhibited /

		Motivational Framing	Motivational Framing		Partial / No — not exhibited
AS	Q36	S7 — Behavioral Appropriateness: Motivational Framing	B: S7 Motivational Framing	Ordinal (1–4)	1–4 scale or NA
AT	Q37	S7 — Revision Recommendations: Motivational Framing	B: S7 Motivational Framing	Open text	Optional
SECTION C: GLOBAL BEHAVIORAL RATINGS					
AU	Q38_1	Global Rating — B1: Diagnostic Questioning	C: Global Ratings	Ordinal (1–4)	1 = Not appropriate, 2 = Somewhat appropriate, 3 = Mostly appropriate, 4 = Fully appropriate
AV	Q38_2	Global Rating — B2: Scaffolded Hint Ladder	C: Global Ratings	Ordinal (1–4)	Same 1–4 scale
AW	Q38_3	Global Rating — B3: Worked Examples with Fading	C: Global Ratings	Ordinal (1–4)	Same 1–4 scale
AX	Q38_4	Global Rating — B4: Mastery Check Gating	C: Global Ratings	Ordinal (1–4)	Same 1–4 scale
AY	Q38_5	Global Rating — B5: Spaced Retrieval Prompts	C: Global Ratings	Ordinal (1–4)	Same 1–4 scale

AZ	Q38_6	Global Rating — B6: Error Classification & Remediation	C: Global Ratings	Ordinal (1–4)	Same 1–4 scale
BA	Q38_7	Global Rating — B7: Motivational Framing	C: Global Ratings	Ordinal (1–4)	Same 1–4 scale
BB	Q39	Global Comments — Cross-cutting patterns, strengths, or concerns across behaviors	C: Global Comments	Open text	Optional
SECTION D: OVERALL DESIGN FIDELITY ASSESSMENT					
BC	Q40	D1 — Overall Design Fidelity Rating	D: Overall Fidelity	Ordinal (1–4)	1 = Low fidelity (<3 behaviors) / 2 = Partial (3–4 behaviors) / 3 = Moderate (5–6 behaviors) / 4 = High fidelity (all 7 behaviors exhibited & appropriate)
BD	Q41	D2 — Primary Strengths: Most consistent design features	D: Overall Fidelity	Open text	"What design features or implemented behaviors did you find most consistent with the theoretical literature?"
BE	Q42	D3 — Primary Weaknesses: Design gaps or misalignments	D: Overall Fidelity	Open text	"What design gaps, inconsistencies, or theoretical

					misalignments were most notable?"
BF	Q43	D4 — Priority Revision Recommendations (top 3 design changes)	D: Overall Fidelity	Open text	"If you could identify three specific design changes that would most improve alignment with theoretical basis"
QUALTRICS SYSTEM FIELDS					
BG	Q_RecaptchaStatus	reCAPTCHA validation status	System	Text	"complete" for all responses
BH	Q_RecaptchaError	reCAPTCHA error details (if any)	System	Text	Blank for all responses (no errors)
THE 7 EVALUATED TUTOR BEHAVIORS — QUICK REFERENCE					
#	Behavior Code	Behavior Name	Description	Scenario Columns	Global Rating Column
1	B1 / S1	Diagnostic Questioning	Tutor asks targeted questions to identify knowledge gaps and misconceptions before	Z, AA, AB (Q11–Q13)	AU (Q38_1)

			instruction		
2	B2 / S2	Scaffolded Hint Ladder	Tutor provides graduated hints from general to specific, avoiding giving away the answer	AC, AD, AE (Q15–Q17)	AV (Q38_2)
3	B3 / S3	Worked Examples with Fading	Tutor demonstrates a complete solution, then progressively removes support on similar problems	AF, AG, AH (Q19–Q21)	AW (Q38_3)
4	B4 / S4	Mastery Check Gating	Tutor requires demonstrated mastery (accuracy + confidence) before advancing to new topics	AI, AJ, AK (Q23–Q25)	AX (Q38_4)
5	B5 / S5	Spaced Retrieval Prompts	Tutor revisits previously learned concepts at intervals to reinforce long-	AL, AM, AN (Q27–Q29)	AY (Q38_5)

			term retention		
6	B6 / S6	Error Classification & Remediation	Tutor distinguishes between procedural and conceptual errors and provides targeted corrective feedback	AO, AP, AQ (Q31–Q33)	AZ (Q38_6)
7	B7 / S7	Motivational Framing	Tutor uses growth-oriented language, normalizes struggle, and reframes errors as learning opportunities	AR, AS, AT (Q35–Q37)	BA (Q38_7)
RATING SCALES REFERENCE					
Behavioral Presence (Section B)					
Yes — fully exhibited	The behavior is clearly and consistently demonstrated				
Partial — exhibited	Some evidence of the behavior, but with gaps or inconsistencies				

but incomplete or inconsistent	
No — not exhibited	The behavior was not observed in the tutor's interactions
Behavioral Appropriateness (Section B) & Global Ratings (Section C)	
4 — Fully appropriate	Strongly consistent with research base; aligns with theoretical expectations
3 — Mostly appropriate	Largely consistent; minor gaps
2 — Somewhat appropriate	Partially consistent; notable gaps between theory and implementation
1 — Not appropriate	Seriously inconsistent with theoretical basis
NA	Behavior not present; select if Presence = No
Overall Design Fidelity (Section D, Q40)	

4 — High fidelity	All 7 behaviors exhibited and appropriately implemented
3 — Moderate fidelity	5–6 behaviors exhibited; minor gaps
2 — Partial fidelity	3–4 behaviors exhibited; some notable weaknesses
1 — Low fidelity	Fewer than 3 behaviors fully or partially exhibited

Part 2: Individual Reviewer Responses

The following pages present the complete response set for each of the eight reviewers (R1 through R8), in the order responses were recorded. Each page contains one reviewer's answers to every survey item, organized by section. Reviewer identifiers (R1-R8) correspond to the same numbering used throughout the main report and supporting analytical sheets.

Reviewer R1

Response Metadata

Start Date

2026-03-31 08:13:19

End Date

2026-03-31 11:46:00

Response Type

IP Address

Progress

100

Duration (in seconds)

12761

Finished

True

Recorded Date

2026-03-31 11:46:01.515000

Response ID

R_1x1TxKm6Q1QLqiB

Section A: Consent and Reviewer Demographics

Do you agree to participate in this study?

I Agree — I have read the informed consent information and agree to participate

Professional Role / Title (Optional)

Higher Education Staff

Institution / Organization

College of Western Idaho

Primary Area of Expertise (select all that apply)

Instructional Design,Learning Science,Community College Education

Years of Relevant Professional Experience

30

Familiarity with LLM-Based Educational Tools

Some familiarity (have used, but not designed)

Conflict of Interest Disclosure (optional) — Any relationship to the researcher or CWI that could bias your evaluation? If none, leave blank.

yes - we are co-workers

Section B - S1: Diagnostic Questioning

S1 — Behavioral Presence: Is the Diagnostic Questioning behavior exhibited by the tutor?

Partial — exhibited but incomplete or inconsistent

S1 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

3 — Mostly appropriate (largely consistent; minor gaps)

S1 — Revision Recommendations: What should change, and why? (Optional)

When the evaluation of the answer is provided a couple of questions are posted but there is no way to respond. Quick strategy check In one short sentence: what steps did you use to solve this? It may be helpful when a wrong answer is provided to as to re-work step by step

Section B - S2: Scaffolded Hint Ladder

S2 — Behavioral Presence: Is the Scaffolded Hint Ladder behavior exhibited by the tutor?

Yes — fully exhibited

S2 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S2 — Revision Recommendations: What should change, and why? (Optional)

I admit. I thought I needed to do the survey last after I did scenarios with the math tutor. I thought the scenarios were to pick different topics to complete.

Section B - S3: Worked Examples with Fading

S3 — Behavioral Presence: Is the Worked Examples with Fading behavior exhibited by the tutor?

Yes — fully exhibited

S3 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S3 — Revision Recommendations: What should change, and why? (Optional)

At one point I was not sure how to set up the problem provided so I ask to see and example of a similar problem and one was provided with an explanation

Section B - S4: Mastery Check Gating

S4 — Behavioral Presence: Is the Mastery Check Gating behavior exhibited by the tutor?

Yes — fully exhibited

S4 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S4 — Revision Recommendations: What should change, and why? (Optional)

I put guessing for my confidence level for my wrong answers. I noticed the coaching was more structured for other answers were I was not guessing. The program asked me to stop and double check even when I had entered the correct answer. That made me think the answer may not be correct but I knew it was, so that made me less confident.

Section B - S5: Spaced Retrieval Prompts

S5 — Behavioral Presence: Is the Spaced Retrieval Prompts behavior exhibited by the tutor?

No — not exhibited

S5 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

[No response]

S5 — Revision Recommendations: What should change, and why? (Optional)

I went from decimals to fractions. During the fraction tutorials not decimal problems were presented

Section B - S6: Error Classification and Remediation

S6 — Behavioral Presence: Is the Error Classification & Remediation behavior exhibited by the tutor?

Yes — fully exhibited

S6 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S6 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S7: Motivational Framing

S7 — Behavioral Presence: Is the Motivational Framing behavior exhibited by the tutor?

Yes — fully exhibited

S7 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S7 — Revision Recommendations: What should change, and why? (Optional)

The responses encourage honesty, understanding, and verifying the correct answer.

Section C: Global Behavioral Ratings

Global Rating - B1: Diagnostic Questioning

4 — Fully appropriate

Global Rating - B2: Scaffolded Hint Ladder

4 — Fully appropriate

Global Rating - B3: Worked Examples with Fading

3 — Mostly appropriate

Global Rating - B4: Mastery Check Gating

4 — Fully appropriate

Global Rating - B5: Spaced Retrieval Prompts

2 — Somewhat appropriate

Global Rating - B6: Error Classification and Remediation

4 — Fully appropriate

Global Rating - B7: Motivational Framing

4 — Fully appropriate

Global Comments - Cross-cutting patterns, strengths, or concerns

I did not see the spaced retrieval but I may have not completed the scenario correctly. It really hung onto the response of guessing long after I did not guess. Sometimes questions are asked that don't provide a specific space for an answer.

Section D: Overall Design Fidelity Assessment

D1 — Overall Design Fidelity Rating: Considering all seven behaviors in aggregate, how would you rate the overall design fidelity of the Math Tutor GPT relative to its behavior specification?

3 — Moderate fidelity (5–6 behaviors exhibited; minor gaps)

D2 — Primary Strengths: What design features or implemented behaviors did you find most consistent with the theoretical literature? Be specific.

Excellent tool for learning and reviewing math concepts. One time I had goal to review and one time to learn. The prompts and initial questions fit both scenario well.

D3 — Primary Weaknesses: What design gaps, inconsistencies, or theoretical misalignments were most notable? Be specific.

If there is not option to provide an answer or if it is meant to be rhetorical then make that more clear. Specifically quick strategy checks.

D4 — Priority Revision Recommendations: If you could identify three specific design changes that would most improve the tutor's alignment with its theoretical basis, what would they be?

See may response in D3

Reviewer R2

Response Metadata

Start Date

2026-04-02 16:47:51

End Date

2026-04-02 18:33:53

Response Type

IP Address

Progress

100

Duration (in seconds)

6362

Finished

True

Recorded Date

2026-04-02 18:33:54.408000

Response ID

R_6wH63a3Dgdvr7Vb

Section A: Consent and Reviewer Demographics

Do you agree to participate in this study?

I Agree — I have read the informed consent information and agree to participate

Professional Role / Title (Optional)

Math Teaching Faculty

Institution / Organization

College of Western Idaho

Primary Area of Expertise (select all that apply)

Mathematics Education,Remedial Mathematics,Community College Education

Years of Relevant Professional Experience

10

Familiarity with LLM-Based Educational Tools

Some familiarity (have used, but not designed)

Conflict of Interest Disclosure (optional) — Any relationship to the researcher or CWI that could bias your evaluation? If none, leave blank.

[No response]

Section B - S1: Diagnostic Questioning

S1 — Behavioral Presence: Is the Diagnostic Questioning behavior exhibited by the tutor?

Yes — fully exhibited

S1 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S1 — Revision Recommendations: What should change, and why? (Optional)

The only thing I would change is the hints or suggested answer format sections. As it asked me to convert a mixed number (2 1/3) to an improper fraction, it would say write your answer like this: 7/3 When it asked me to articulate my process for solving it would also provide the correct answer as a suggestion. For example, it told me to explain my steps for converting a mixed number and then it recommended that i write something like multiply the whole number by the denominator and add the numerator. For multiple problems it gave me the answer before I submitted my own response.

Section B - S2: Scaffolded Hint Ladder

S2 — Behavioral Presence: Is the Scaffolded Hint Ladder behavior exhibited by the tutor?

Yes — fully exhibited

S2 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S2 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S3: Worked Examples with Fading

S3 — Behavioral Presence: Is the Worked Examples with Fading behavior exhibited by the tutor?

Yes — fully exhibited

S3 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S3 — Revision Recommendations: What should change, and why? (Optional)

I like that it adds more details if I get a problem wrong. It will have me just focus on the first step or will have me follow an example with baked in guidance. As my confidence grows, it slowly removes those crutches.

Section B - S4: Mastery Check Gating

S4 — Behavioral Presence: Is the Mastery Check Gating behavior exhibited by the tutor?

Yes — fully exhibited

S4 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S4 — Revision Recommendations: What should change, and why? (Optional)

When it asked how I got my answer i told it that I used a calculator. To which it responded "Using a calculator is fine in real life, but for the placement test you usually won’t have one, so we want you to be able to do these mentally or on scratch paper." This is not true, students are allowed to use a calculator on the test.

Section B - S5: Spaced Retrieval Prompts

S5 — Behavioral Presence: Is the Spaced Retrieval Prompts behavior exhibited by the tutor?

Partial — exhibited but incomplete or inconsistent

S5 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

3 — Mostly appropriate (largely consistent; minor gaps)

S5 — Revision Recommendations: What should change, and why? (Optional)

ChatGPT told me that I reached my message limit and I need to pay to continue. I had mastered solving equations, then asked to practice multiplying decimals. I was not given a review problem while learning how to multiply decimals. After the waiting period was up and I mastered the decimals, then it returned to the solving equations concept. I feel like for this feature o work effectively, students need to pay for chatGPT.

Section B - S6: Error Classification and Remediation

S6 — Behavioral Presence: Is the Error Classification & Remediation behavior exhibited by the tutor?

Yes — fully exhibited

S6 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S6 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S7: Motivational Framing

S7 — Behavioral Presence: Is the Motivational Framing behavior exhibited by the tutor?

Yes — fully exhibited

S7 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S7 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section C: Global Behavioral Ratings

Global Rating - B1: Diagnostic Questioning

3 — Mostly appropriate

Global Rating - B2: Scaffolded Hint Ladder

3 — Mostly appropriate

Global Rating - B3: Worked Examples with Fading

4 — Fully appropriate

Global Rating - B4: Mastery Check Gating

4 — Fully appropriate

Global Rating - B5: Spaced Retrieval Prompts

2 — Somewhat appropriate

Global Rating - B6: Error Classification and Remediation

2 — Somewhat appropriate

Global Rating - B7: Motivational Framing

4 — Fully appropriate

Global Comments - Cross-cutting patterns, strengths, or concerns

I usually would run out of prompt before I could reach the spaced retrieval prompt feature. I also felt like it was giving a lot of unprompted leading hints that were just the answer when doing problems with fractions. That said, the more I used it the better it got at each of the behaviors listed above.

Section D: Overall Design Fidelity Assessment

D1 — Overall Design Fidelity Rating: Considering all seven behaviors in aggregate, how would you rate the overall design fidelity of the Math Tutor GPT relative to its behavior specification?

3 — Moderate fidelity (5–6 behaviors exhibited; minor gaps)

D2 — Primary Strengths: What design features or implemented behaviors did you find most consistent with the theoretical literature? Be specific.

mastery check gating - it made sure I was confident and accurate in my answers before raising the difficulty, diagnostic questioning - the more I expressed my level of understanding the better it was at giving me questions based on my needs (confidence building for low skills vs testing mastery for high skills) , self explanation prompting - it made me do this ALOT my only issue is that it would normally tell me what to write after asking me to articulate my steps.

D3 — Primary Weaknesses: What design gaps, inconsistencies, or theoretical misalignments were most notable? Be specific.

Scaffolded hint ladder - it's pretty eager to fully explain the concept from the get go. that said it does a great job removing those supports as I got more questions correct. The exception to this was the fraction problems. it would say write your answer in this form: then it would just give me the answer before I got to try it on my own.

D4 — Priority Revision Recommendations: If you could identify three specific design changes that would most improve the tutor's alignment with its theoretical basis, what would they be?

Allow students to submit more prompts without having to pay for CHAT GPT. Ensure that the hints don't solve the problem for you. When you ask a student to explain their process, don't tell them what to write.

Reviewer R3

Response Metadata

Start Date

2026-04-06 14:00:09

End Date

2026-04-06 14:05:40

Response Type

IP Address

Progress

100

Duration (in seconds)

331

Finished

True

Recorded Date

2026-04-06 14:05:41.076000

Response ID

R_1SK0W3mBnVJRl85

Section A: Consent and Reviewer Demographics

Do you agree to participate in this study?

I Agree — I have read the informed consent information and agree to participate

Professional Role / Title (Optional)

Instructional Designer

Institution / Organization

College of Western Idaho

Primary Area of Expertise (select all that apply)

Instructional Design

Years of Relevant Professional Experience

4

Familiarity with LLM-Based Educational Tools

Moderate (have designed or evaluated LLM tools)

Conflict of Interest Disclosure (optional) — Any relationship to the researcher or CWI that could bias your evaluation? If none, leave blank.

[No response]

Section B - S1: Diagnostic Questioning

S1 — Behavioral Presence: Is the Diagnostic Questioning behavior exhibited by the tutor?

Yes — fully exhibited

S1 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S1 — Revision Recommendations: What should change, and why? (Optional)

None at this time.

Section B - S2: Scaffolded Hint Ladder

S2 — Behavioral Presence: Is the Scaffolded Hint Ladder behavior exhibited by the tutor?

Yes — fully exhibited

S2 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S2 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S3: Worked Examples with Fading

S3 — Behavioral Presence: Is the Worked Examples with Fading behavior exhibited by the tutor?

Yes — fully exhibited

S3 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S3 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S4: Mastery Check Gating

S4 — Behavioral Presence: Is the Mastery Check Gating behavior exhibited by the tutor?

Yes — fully exhibited

S4 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S4 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S5: Spaced Retrieval Prompts

S5 — Behavioral Presence: Is the Spaced Retrieval Prompts behavior exhibited by the tutor?

Partial — exhibited but incomplete or inconsistent

S5 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

3 — Mostly appropriate (largely consistent; minor gaps)

S5 — Revision Recommendations: What should change, and why? (Optional)

I noticed sometimes there is a delay between responses.

Section B - S6: Error Classification and Remediation

S6 — Behavioral Presence: Is the Error Classification & Remediation behavior exhibited by the tutor?

Yes — fully exhibited

S6 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S6 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S7: Motivational Framing

S7 — Behavioral Presence: Is the Motivational Framing behavior exhibited by the tutor?

Yes — fully exhibited

S7 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S7 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section C: Global Behavioral Ratings

Global Rating - B1: Diagnostic Questioning

4 — Fully appropriate

Global Rating - B2: Scaffolded Hint Ladder

4 — Fully appropriate

Global Rating - B3: Worked Examples with Fading

4 — Fully appropriate

Global Rating - B4: Mastery Check Gating

4 — Fully appropriate

Global Rating - B5: Spaced Retrieval Prompts

3 — Mostly appropriate

Global Rating - B6: Error Classification and Remediation

3 — Mostly appropriate

Global Rating - B7: Motivational Framing

4 — Fully appropriate

Global Comments - Cross-cutting patterns, strengths, or concerns

There was one time where I had it correct and it almost marked it as incorrect, but then corrected itself.

Section D: Overall Design Fidelity Assessment

D1 — Overall Design Fidelity Rating: Considering all seven behaviors in aggregate, how would you rate the overall design fidelity of the Math Tutor GPT relative to its behavior specification?

4 — High fidelity (all 7 behaviors exhibited and appropriately implemented)

D2 — Primary Strengths: What design features or implemented behaviors did you find most consistent with the theoretical literature? Be specific.

The prompting features for when it corrected my responses were spot on.

D3 — Primary Weaknesses: What design gaps, inconsistencies, or theoretical misalignments were most notable? Be specific.

Slight delay when processing immediate feedback.

D4 — Priority Revision Recommendations: If you could identify three specific design changes that would most improve the tutor's alignment with its theoretical basis, what would they be?

None at this time.

Reviewer R4

Response Metadata

Start Date

2026-04-07 06:26:45

End Date

2026-04-07 08:29:14

Response Type

IP Address

Progress

100

Duration (in seconds)

7348

Finished

True

Recorded Date

2026-04-07 08:29:15.162000

Response ID

R_3dNceEFNxYRFnqT

Section A: Consent and Reviewer Demographics

Do you agree to participate in this study?

I Agree — I have read the informed consent information and agree to participate

Professional Role / Title (Optional)

Instructor

Institution / Organization

College of Western Idaho

Primary Area of Expertise (select all that apply)

Instructional Design,Mathematics Education,Community College Education,Artificial Intelligence / LLM Tool Design

Years of Relevant Professional Experience

3

Familiarity with LLM-Based Educational Tools

Extensive (research or development background)

Conflict of Interest Disclosure (optional) — Any relationship to the researcher or CWI that could bias your evaluation? If none, leave blank.

[No response]

Section B - S1: Diagnostic Questioning

S1 — Behavioral Presence: Is the Diagnostic Questioning behavior exhibited by the tutor?

Partial — exhibited but incomplete or inconsistent

S1 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

3 — Mostly appropriate (largely consistent; minor gaps)

S1 — Revision Recommendations: What should change, and why? (Optional)

The tutor should verbalize routing decisions (e.g., "You've shown solid simplification skills, so I'm moving to mixed numbers"). Also worth flagging: the tutor initially confirmed an incorrect answer as correct before reversing via a confidence probe. A student who answers "very sure" may not get the correction. The confidence probe mechanism is useful, but it shouldn't be the only safeguard against a misgraded response.

Section B - S2: Scaffolded Hint Ladder

S2 — Behavioral Presence: Is the Scaffolded Hint Ladder behavior exhibited by the tutor?

Yes — fully exhibited

S2 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

3 — Mostly appropriate (largely consistent; minor gaps)

S2 — Revision Recommendations: What should change, and why? (Optional)

Hints may be overly comprehensive on first step. Presenting conceptual questions prior to equation structure may assist in building concept understanding.

Section B - S3: Worked Examples with Fading

S3 — Behavioral Presence: Is the Worked Examples with Fading behavior exhibited by the tutor?

Partial — exhibited but incomplete or inconsistent

S3 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

2 — Somewhat appropriate (partially consistent; notable gaps)

S3 — Revision Recommendations: What should change, and why? (Optional)

Incorrect answer resulted in return of full solution rather than hints. Transition between stages are not explicitly named (e.g. "now try a simpler version" or "now try without any hints"), which would help the student understand the flow of the practice.

Section B - S4: Mastery Check Gating

S4 — Behavioral Presence: Is the Mastery Check Gating behavior exhibited by the tutor?

Yes — fully exhibited

S4 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

2 — Somewhat appropriate (partially consistent; notable gaps)

S4 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S5: Spaced Retrieval Prompts

S5 — Behavioral Presence: Is the Spaced Retrieval Prompts behavior exhibited by the tutor?

No — not exhibited

S5 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

NA — Behavior not present (select if Presence = No)

S5 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S6: Error Classification and Remediation

S6 — Behavioral Presence: Is the Error Classification & Remediation behavior exhibited by the tutor?

Partial — exhibited but incomplete or inconsistent

S6 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

2 — Somewhat appropriate (partially consistent; notable gaps)

S6 — Revision Recommendations: What should change, and why? (Optional)

For conceptual errors: address the underlying misunderstanding (e.g., "Your answer of 3/8 suggests you added the denominators")

Section B - S7: Motivational Framing

S7 — Behavioral Presence: Is the Motivational Framing behavior exhibited by the tutor?

Yes — fully exhibited

S7 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S7 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section C: Global Behavioral Ratings

Global Rating - B1: Diagnostic Questioning

3 — Mostly appropriate

Global Rating - B2: Scaffolded Hint Ladder

2 — Somewhat appropriate

Global Rating - B3: Worked Examples with Fading

2 — Somewhat appropriate

Global Rating - B4: Mastery Check Gating

2 — Somewhat appropriate

Global Rating - B5: Spaced Retrieval Prompts

1 — Not appropriate

Global Rating - B6: Error Classification and Remediation

2 — Somewhat appropriate

Global Rating - B7: Motivational Framing

3 — Mostly appropriate

Global Comments - Cross-cutting patterns, strengths, or concerns

Motivational framing and diagnostic opening are genuine strengths.

Section D: Overall Design Fidelity Assessment

D1 — Overall Design Fidelity Rating: Considering all seven behaviors in aggregate, how would you rate the overall design fidelity of the Math Tutor GPT relative to its behavior specification?

3 — Moderate fidelity (5–6 behaviors exhibited; minor gaps)

D2 — Primary Strengths: What design features or implemented behaviors did you find most consistent with the theoretical literature? Be specific.

Motivational Framing is the standout. Explicit plan/execution separation on errors, and direct growth-oriented reframing of negative self-talk are great. The confidence probe ("guessing, somewhat sure, very sure") integrated throughout every session is also a nice touch.

D3 — Primary Weaknesses: What design gaps, inconsistencies, or theoretical misalignments were most notable? Be specific.

Spaced retrieval is absent. procedural and conceptual errors receive the same corrective pattern, which misses the theoretical core of the behavior. Hints sometimes skip directly to the answer, and examples don't progressively reduce support toward independence.

D4 — Priority Revision Recommendations: If you could identify three specific design changes that would most improve the tutor's alignment with its theoretical basis, what would they be?

1. Spaced retrieval mechanism 2. Differentiation on error responses based on error type 3. Multi-tier hint ladder.

Reviewer R5

Response Metadata

Start Date

2026-04-08 09:51:50

End Date

2026-04-08 13:42:22

Response Type

IP Address

Progress

100

Duration (in seconds)

13831

Finished

True

Recorded Date

2026-04-08 13:42:22.770000

Response ID

R_6sR8k1TmoAlRjFw

Section A: Consent and Reviewer Demographics

Do you agree to participate in this study?

I Agree — I have read the informed consent information and agree to participate

Professional Role / Title (Optional)

[No response]

Institution / Organization

College of Western Idaho

Primary Area of Expertise (select all that apply)

Instructional Design

Years of Relevant Professional Experience

17

Familiarity with LLM-Based Educational Tools

Some familiarity (have used, but not designed)

Conflict of Interest Disclosure (optional) — Any relationship to the researcher or CWI that could bias your evaluation? If none, leave blank.

[No response]

Section B - S1: Diagnostic Questioning

S1 — Behavioral Presence: Is the Diagnostic Questioning behavior exhibited by the tutor?

Yes — fully exhibited

S1 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

3 — Mostly appropriate (largely consistent; minor gaps)

S1 — Revision Recommendations: What should change, and why? (Optional)

As far as I can tell, it did not explain why it was starting at a particular content level.

Section B - S2: Scaffolded Hint Ladder

S2 — Behavioral Presence: Is the Scaffolded Hint Ladder behavior exhibited by the tutor?

Yes — fully exhibited

S2 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S2 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S3: Worked Examples with Fading

S3 — Behavioral Presence: Is the Worked Examples with Fading behavior exhibited by the tutor?

Yes — fully exhibited

S3 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S3 — Revision Recommendations: What should change, and why? (Optional)

one time, I got the answer correct; it said" not quite", worked the problem, and then said So, [the answer] is correct. The question was: Try this one: 6 x + 1 = 19 6x+1=19 What is x x? I said 3 It said: Not quite. Start with 6 x + 1 = 19 6x+1=19 Undo the + 1 +1 first: 6 x + 1 = 19 6 x = 18 6x+1 6x =19 =18 Then divide by 6 6: 6 x = 18 x = 3 6x x =18 =3 So x = 3 x=3 is correct. Nice.

Section B - S4: Mastery Check Gating

S4 — Behavioral Presence: Is the Mastery Check Gating behavior exhibited by the tutor?

Partial — exhibited but incomplete or inconsistent

S4 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

3 — Mostly appropriate (largely consistent; minor gaps)

S4 — Revision Recommendations: What should change, and why? (Optional)

None

Section B - S5: Spaced Retrieval Prompts

S5 — Behavioral Presence: Is the Spaced Retrieval Prompts behavior exhibited by the tutor?

No — not exhibited

S5 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

[No response]

S5 — Revision Recommendations: What should change, and why? (Optional)

ISSUES: I got a question correct. It told me not quite and then that I was correct. Conversation: Now one small change: x+6=16 Send just the number. 10 Not quite. x+6=16 means □+6=16 Ask: What number plus 6 equals 16? Check it: 10+6=16 So: x=10 ------- Then I answered wrong, it told me "Correct" but then said the check didn't match the problem. It asked me to try the problem again. I got it right, it told me that on the last page it made a mistake and I also got it right, then. However, it was not correct. Conversation: What number plus 4 equals 15? Send just the number. 9 Correct. 9+4=13 Sorry — that check does not match the problem. The problem was: x+4=15 So we need: 9+4=13 That is too small. Try again: □+4=15 What number goes in the box? Send just the number. Correct. 11+4=15 So: x=11 And yes — your last answer was right. I made a checking mistake in the previous message.

Section B - S6: Error Classification and Remediation

S6 — Behavioral Presence: Is the Error Classification & Remediation behavior exhibited by the tutor?

Partial — exhibited but incomplete or inconsistent

S6 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

2 — Somewhat appropriate (partially consistent; notable gaps)

S6 — Revision Recommendations: What should change, and why? (Optional)

I kept getting the answer wrong for different reasons. Many times it didn't give me any help/remediation (I did not ask for help.)

Section B - S7: Motivational Framing

S7 — Behavioral Presence: Is the Motivational Framing behavior exhibited by the tutor?

Yes — fully exhibited

S7 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S7 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section C: Global Behavioral Ratings

Global Rating - B1: Diagnostic Questioning

3 — Mostly appropriate

Global Rating - B2: Scaffolded Hint Ladder

2 — Somewhat appropriate

Global Rating - B3: Worked Examples with Fading

4 — Fully appropriate

Global Rating - B4: Mastery Check Gating

4 — Fully appropriate

Global Rating - B5: Spaced Retrieval Prompts

2 — Somewhat appropriate

Global Rating - B6: Error Classification and Remediation

3 — Mostly appropriate

Global Rating - B7: Motivational Framing

3 — Mostly appropriate

Global Comments - Cross-cutting patterns, strengths, or concerns

[No response]

Section D: Overall Design Fidelity Assessment

D1 — Overall Design Fidelity Rating: Considering all seven behaviors in aggregate, how would you rate the overall design fidelity of the Math Tutor GPT relative to its behavior specification?

3 — Moderate fidelity (5–6 behaviors exhibited; minor gaps)

D2 — Primary Strengths: What design features or implemented behaviors did you find most consistent with the theoretical literature? Be specific.

[No response]

D3 — Primary Weaknesses: What design gaps, inconsistencies, or theoretical misalignments were most notable? Be specific.

* I felt it could have been friendlier. It was too robotic. Same sort of two-word encouragements. *I think that when it said you were correct or incorrect, it should start with "You said [the answer I gave]." There was one time when I answered at least 15 times incorrectly in a row. It did not offer any escalated assistance and did not go all the way back to a simpler math concept. I purposely didn't ask it for help as the student might not ask. There should be a threshold for wrong answers and then at least ask if you need help or hints.

D4 — Priority Revision Recommendations: If you could identify three specific design changes that would most improve the tutor's alignment with its theoretical basis, what would they be?

* I mentioned in one answer that the feedback/info about what was right/wrong was the wrong feedback. *There should be a threshold for wrong answers and then at least ask if you need help or hints. *I felt it could have been friendlier. It was too robotic. Same sort of two-word encouragements.Needs to be more conversational, otherwise I think it increases a "this is boring" factor.

Reviewer R6

Response Metadata

Start Date

2026-04-08 16:00:30

End Date

2026-04-08 16:17:16

Response Type

IP Address

Progress

100

Duration (in seconds)

1005

Finished

True

Recorded Date

2026-04-08 16:17:17.035000

Response ID

R_7mBtMZqmYb9Whc5

Section A: Consent and Reviewer Demographics

Do you agree to participate in this study?

I Agree — I have read the informed consent information and agree to participate

Professional Role / Title (Optional)

[No response]

Institution / Organization

CWI

Primary Area of Expertise (select all that apply)

Mathematics Education

Years of Relevant Professional Experience

15

Familiarity with LLM-Based Educational Tools

No prior experience

Conflict of Interest Disclosure (optional) — Any relationship to the researcher or CWI that could bias your evaluation? If none, leave blank.

I am a math professor at CWI.

Section B - S1: Diagnostic Questioning

S1 — Behavioral Presence: Is the Diagnostic Questioning behavior exhibited by the tutor?

Partial — exhibited but incomplete or inconsistent

S1 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

[No response]

S1 — Revision Recommendations: What should change, and why? (Optional)

What is diagnostic questioning behavior? Don't understand the questions being asked. Feedback uses math terms students don't understand. Couldn't figure out how to answer questions.

Section B - S2: Scaffolded Hint Ladder

S2 — Behavioral Presence: Is the Scaffolded Hint Ladder behavior exhibited by the tutor?

Yes — fully exhibited

S2 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

2 — Somewhat appropriate (partially consistent; notable gaps)

S2 — Revision Recommendations: What should change, and why? (Optional)

Don't undrstand technical jargon enough to answer the survey questions. The hint is procedural not conceptual.

Section B - S3: Worked Examples with Fading

S3 — Behavioral Presence: Is the Worked Examples with Fading behavior exhibited by the tutor?

No — not exhibited

S3 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

1 — Not appropriate (seriously inconsistent with theoretical basis)

S3 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S4: Mastery Check Gating

S4 — Behavioral Presence: Is the Mastery Check Gating behavior exhibited by the tutor?

[No response]

S4 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

[No response]

S4 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S5: Spaced Retrieval Prompts

S5 — Behavioral Presence: Is the Spaced Retrieval Prompts behavior exhibited by the tutor?

[No response]

S5 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

[No response]

S5 — Revision Recommendations: What should change, and why? (Optional)

don't have time to do this one.

Section B - S6: Error Classification and Remediation

S6 — Behavioral Presence: Is the Error Classification & Remediation behavior exhibited by the tutor?

[No response]

S6 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

[No response]

S6 — Revision Recommendations: What should change, and why? (Optional)

I didnt see anything about identifying procedural and conceptual errors.

Section B - S7: Motivational Framing

S7 — Behavioral Presence: Is the Motivational Framing behavior exhibited by the tutor?

[No response]

S7 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

[No response]

S7 — Revision Recommendations: What should change, and why? (Optional)

language seemed OK. Kept using terms student won't understand.

Section C: Global Behavioral Ratings

Global Rating - B1: Diagnostic Questioning

2 — Somewhat appropriate

Global Rating - B2: Scaffolded Hint Ladder

2 — Somewhat appropriate

Global Rating - B3: Worked Examples with Fading

2 — Somewhat appropriate

Global Rating - B4: Mastery Check Gating

2 — Somewhat appropriate

Global Rating - B5: Spaced Retrieval Prompts

[No response]

Global Rating - B6: Error Classification and Remediation

1 — Not appropriate

Global Rating - B7: Motivational Framing

2 — Somewhat appropriate

Global Comments - Cross-cutting patterns, strengths, or concerns

[No response]

Section D: Overall Design Fidelity Assessment

D1 — Overall Design Fidelity Rating: Considering all seven behaviors in aggregate, how would you rate the overall design fidelity of the Math Tutor GPT relative to its behavior specification?

1 — Low fidelity (fewer than 3 behaviors fully or partially exhibited)

D2 — Primary Strengths: What design features or implemented behaviors did you find most consistent with the theoretical literature? Be specific.

[No response]

D3 — Primary Weaknesses: What design gaps, inconsistencies, or theoretical misalignments were most notable? Be specific.

Hints, explanations, use terms students won't understand. No distinguishing between conceptual and procedural. I couldn't figure out how I am supposed to enter answers.

D4 — Priority Revision Recommendations: If you could identify three specific design changes that would most improve the tutor's alignment with its theoretical basis, what would they be?

[No response]

Reviewer R7

Response Metadata

Start Date

2026-04-08 17:32:55

End Date

2026-04-08 18:01:00

Response Type

IP Address

Progress

100

Duration (in seconds)

1685

Finished

True

Recorded Date

2026-04-08 18:01:01.333000

Response ID

R_7MPXArApdkdmbTf

Section A: Consent and Reviewer Demographics

Do you agree to participate in this study?

I Agree — I have read the informed consent information and agree to participate

Professional Role / Title (Optional)

Math Instructor, College of Western Idaho

Institution / Organization

College of Western Idaho

Primary Area of Expertise (select all that apply)

Mathematics Education,Community College Education,Student Success,Artificial Intelligence / LLM Tool Design

Years of Relevant Professional Experience

20 or so

Familiarity with LLM-Based Educational Tools

Extensive (research or development background)

Conflict of Interest Disclosure (optional) — Any relationship to the researcher or CWI that could bias your evaluation? If none, leave blank.

I know Andrew personally and respect his work, but my opinion of him has not affected my evaluation of this tool.

Section B - S1: Diagnostic Questioning

S1 — Behavioral Presence: Is the Diagnostic Questioning behavior exhibited by the tutor?

Yes — fully exhibited

S1 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S1 — Revision Recommendations: What should change, and why? (Optional)

The tool did a good job of asking where I felt I was mathematically, but at each subsequent step, it asked for my confidence in my answer but then didn't pause for an response and just went ahead with its explanation of my (plentiful) mistakes.

Section B - S2: Scaffolded Hint Ladder

S2 — Behavioral Presence: Is the Scaffolded Hint Ladder behavior exhibited by the tutor?

Yes — fully exhibited

S2 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S2 — Revision Recommendations: What should change, and why? (Optional)

I begged and pleaded to just get the answer, promising the AI candy and even strawberries, but it refused to budge, and at most gave me only the next step of the process, never the final answer.

Section B - S3: Worked Examples with Fading

S3 — Behavioral Presence: Is the Worked Examples with Fading behavior exhibited by the tutor?

Yes — fully exhibited

S3 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S3 — Revision Recommendations: What should change, and why? (Optional)

I tried all kinds of typical mistakes on this one and the bot did a good job of navigating through them and giving me appropriate advice. It also seemed to know what sort of mistake and made that led to the incorrect answer. I never reduced my fractions after dividing by the x coefficient, and it never called me on it except in one case, where it wanted me to write the answer as a mixed number instead of an improper fraction (all the other fractions would reduce to integers and it seemed happy to do that for me, because I never did reduce any of them over five or six problem)s

Section B - S4: Mastery Check Gating

S4 — Behavioral Presence: Is the Mastery Check Gating behavior exhibited by the tutor?

Yes — fully exhibited

S4 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S4 — Revision Recommendations: What should change, and why? (Optional)

I made a lot of mistakes early on, and only gradually started getting correct answers. Later on, when I made a similar mistake on the diagnostic test, it pointed out to me that I had learned this before and asked me to work another example of that type of problem.

Section B - S5: Spaced Retrieval Prompts

S5 — Behavioral Presence: Is the Spaced Retrieval Prompts behavior exhibited by the tutor?

Yes — fully exhibited

S5 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S5 — Revision Recommendations: What should change, and why? (Optional)

See my previous recommendation. The bot seemed to do a good job of representing problems that tested earlier concepts.

Section B - S6: Error Classification and Remediation

S6 — Behavioral Presence: Is the Error Classification & Remediation behavior exhibited by the tutor?

Yes — fully exhibited

S6 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S6 — Revision Recommendations: What should change, and why? (Optional)

It did a good job of recognizing what types of errors I had made. (I have to do the same thing when grading exams, so I'm very aware of whether the bot was correct in this or not.)

Section B - S7: Motivational Framing

S7 — Behavioral Presence: Is the Motivational Framing behavior exhibited by the tutor?

Yes — fully exhibited

S7 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S7 — Revision Recommendations: What should change, and why? (Optional)

The bot was appropriately encouraging and seemed to be relatively free of glazing behavior.

Section C: Global Behavioral Ratings

Global Rating - B1: Diagnostic Questioning

3 — Mostly appropriate

Global Rating - B2: Scaffolded Hint Ladder

3 — Mostly appropriate

Global Rating - B3: Worked Examples with Fading

4 — Fully appropriate

Global Rating - B4: Mastery Check Gating

4 — Fully appropriate

Global Rating - B5: Spaced Retrieval Prompts

4 — Fully appropriate

Global Rating - B6: Error Classification and Remediation

4 — Fully appropriate

Global Rating - B7: Motivational Framing

4 — Fully appropriate

Global Comments - Cross-cutting patterns, strengths, or concerns

The bot would occasionally give me a "Quick Strategy check (1 sentence)" and ask for an explanation of the next step in a process, but never waited for an answer. I'm not sure if a student would actually talk through this without more insistence from the bot, and this would probably benefit the student if they had to.

Section D: Overall Design Fidelity Assessment

D1 — Overall Design Fidelity Rating: Considering all seven behaviors in aggregate, how would you rate the overall design fidelity of the Math Tutor GPT relative to its behavior specification?

4 — High fidelity (all 7 behaviors exhibited and appropriately implemented)

D2 — Primary Strengths: What design features or implemented behaviors did you find most consistent with the theoretical literature? Be specific.

This looks like a good tool for students at this level to improve their math skills. I would like to see more of it.

D3 — Primary Weaknesses: What design gaps, inconsistencies, or theoretical misalignments were most notable? Be specific.

One thing that concerned me was whether a student at this level would be comfortable putting in answers in a correct format. Some students would need help with this, I think.

D4 — Priority Revision Recommendations: If you could identify three specific design changes that would most improve the tutor's alignment with its theoretical basis, what would they be?

It seems well aligned, but to repeat some of my earlier observations, it would be better if the bot would pause for an answer to the questions of confidence in the answer and a one-sentence summary of the next step. Otherwise, this looks like a great tool.

Reviewer R8

Response Metadata

Start Date

2026-04-08 19:10:31

End Date

2026-04-08 20:13:43

Response Type

IP Address

Progress

100

Duration (in seconds)

3791

Finished

True

Recorded Date

2026-04-08 20:13:43.576000

Response ID

R_55xpDzP4E36pgWd

Section A: Consent and Reviewer Demographics

Do you agree to participate in this study?

I Agree — I have read the informed consent information and agree to participate

Professional Role / Title (Optional)

Instructional Designer

Institution / Organization

College of Western Idaho

Primary Area of Expertise (select all that apply)

Instructional Design

Years of Relevant Professional Experience

12+

Familiarity with LLM-Based Educational Tools

Moderate (have designed or evaluated LLM tools)

Conflict of Interest Disclosure (optional) — Any relationship to the researcher or CWI that could bias your evaluation? If none, leave blank.

Co-worker

Section B - S1: Diagnostic Questioning

S1 — Behavioral Presence: Is the Diagnostic Questioning behavior exhibited by the tutor?

Partial — exhibited but incomplete or inconsistent

S1 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

3 — Mostly appropriate (largely consistent; minor gaps)

S1 — Revision Recommendations: What should change, and why? (Optional)

Overall, it may be helpful to add more conversation starters for users to initially view. The conversation starters at the bottom of the placement test skills menu are not very prominent. Will the same procedure be used each time, or, later can the students have more choice in how to use the tutor?

Section B - S2: Scaffolded Hint Ladder

S2 — Behavioral Presence: Is the Scaffolded Hint Ladder behavior exhibited by the tutor?

Yes — fully exhibited

S2 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S2 — Revision Recommendations: What should change, and why? (Optional)

I liked the scaffolded steps, especially the ability to provide a response about the first step to take toward solving the problem and then working with the tutor from there. It was helpful that the tutor asked the user to state their confidence level and then would provide follow up and encouragement based on the level of confidence. In a later scenario, the tutor forgot to ask for an initial answer before giving a hint, so that could be improved to make the user experience a productive struggle.

Section B - S3: Worked Examples with Fading

S3 — Behavioral Presence: Is the Worked Examples with Fading behavior exhibited by the tutor?

Yes — fully exhibited

S3 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S3 — Revision Recommendations: What should change, and why? (Optional)

I like that the tutor provides the fully annotated process with an example. Then it asks you to complete the first step of a new problem. It follows it up by feedback for the first step and then asks you to continue to solve the problem. If the user doesn't know how to proceed, the tutor offers a hint about how to proceed or how to complete the next step, but still expects the user to finish the problem.

Section B - S4: Mastery Check Gating

S4 — Behavioral Presence: Is the Mastery Check Gating behavior exhibited by the tutor?

Partial — exhibited but incomplete or inconsistent

S4 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

3 — Mostly appropriate (largely consistent; minor gaps)

S4 — Revision Recommendations: What should change, and why? (Optional)

I completed several initial problems regarding fractions. It seems like I could have continued in this way infinitely. I would have preferred to get larger picture feedback, especially if the tutor thought I was ready to proceed to the next Unit 0 topic.

Section B - S5: Spaced Retrieval Prompts

S5 — Behavioral Presence: Is the Spaced Retrieval Prompts behavior exhibited by the tutor?

No — not exhibited

S5 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

2 — Somewhat appropriate (partially consistent; notable gaps)

S5 — Revision Recommendations: What should change, and why? (Optional)

I thought I had worked on certain topics long enough for mastery, but was not signaled with spaced retrieval prompts. Maybe this required more time or interactions than I had time for.

Section B - S6: Error Classification and Remediation

S6 — Behavioral Presence: Is the Error Classification & Remediation behavior exhibited by the tutor?

Yes — fully exhibited

S6 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S6 — Revision Recommendations: What should change, and why? (Optional)

[No response]

Section B - S7: Motivational Framing

S7 — Behavioral Presence: Is the Motivational Framing behavior exhibited by the tutor?

Yes — fully exhibited

S7 — Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

4 — Fully appropriate (strongly consistent; aligns with research base)

S7 — Revision Recommendations: What should change, and why? (Optional)

I appreciate that the tutor points out what the learner can do well and indicates areas for growth. The tutor points out behavior patterns such as when a new skill is presented, the learner experiences uncertainty about the first move.

Section C: Global Behavioral Ratings

Global Rating - B1: Diagnostic Questioning

3 — Mostly appropriate

Global Rating - B2: Scaffolded Hint Ladder

3 — Mostly appropriate

Global Rating - B3: Worked Examples with Fading

3 — Mostly appropriate

Global Rating - B4: Mastery Check Gating

2 — Somewhat appropriate

Global Rating - B5: Spaced Retrieval Prompts

2 — Somewhat appropriate

Global Rating - B6: Error Classification and Remediation

4 — Fully appropriate

Global Rating - B7: Motivational Framing

4 — Fully appropriate

Global Comments - Cross-cutting patterns, strengths, or concerns

[No response]

Section D: Overall Design Fidelity Assessment

D1 — Overall Design Fidelity Rating: Considering all seven behaviors in aggregate, how would you rate the overall design fidelity of the Math Tutor GPT relative to its behavior specification?

3 — Moderate fidelity (5–6 behaviors exhibited; minor gaps)

D2 — Primary Strengths: What design features or implemented behaviors did you find most consistent with the theoretical literature? Be specific.

I think the tutor works well within the subject area and is very encouraging in an area where learners feel very vulnerable. I like that the tutor offers an overview about the specific concept and then asks the learn to engage with the first step and then proceeds from there. I also like the check-ins for confidence levels for answers.

D3 — Primary Weaknesses: What design gaps, inconsistencies, or theoretical misalignments were most notable? Be specific.

[No response]

D4 — Priority Revision Recommendations: If you could identify three specific design changes that would most improve the tutor's alignment with its theoretical basis, what would they be?

I would provide more specific conversations starters in the beginning. Or, I would provide more specific information about how to proceed after the tutor outlines the topics for the placement exam. I assume you'll be adding these components for more varied use by students outside of this research trial? I would also include slightly more productive struggle. Sometimes, the tutor is too eager to provide step-by-step instructions about how to solve the problem vs. letting the learner try to initiate with the problem and then seek assistance. Also, I think that the feedback for mastery needs to come sooner in the interactions with learners (or at least, a check-in needs to occur to let the learner know how they are doing with the concept with the potential option to proceed with more practice on the topic or to move on.).

Appendix C: Glossary

The following terms are used throughout this report. Definitions are provided for clarity and to support readers from outside the instructional design or AI-tutoring research communities.

A Priori Code. A code defined before you look at the current data, usually based on theory, prior studies, or your research questions.

AI (Artificial Intelligence). Computer systems designed to perform tasks that typically require human intelligence, such as understanding language, recognizing patterns, or making decisions.

API (Application Programming Interface).

Asynchronous Review. Review of materials, prototypes, or data that occurs at different times rather than in real-time meetings, often via documents, recordings, or online tools.

Behavioral Appropriateness. The extent to which a learner’s or tutor’s observable behaviors fit contextual norms and expectations for professionalism, respect, and task focus.

Behavioral Presence. The degree to which a tutor (human or AI) is perceived as actively engaged through its observable actions, such as timely responses, prompts, and feedback moves.

ChatGPT. An interactive application built on GPT-style LLMs that lets users have conversational back-and-forth in natural language, often with added instruction-tuning and safety layers.

Cognitive Load Theory. A theory that explains how the limited capacity of working memory affects learning, emphasizing that instruction should reduce unnecessary load and support processing of essential material.

Context Window. The maximum amount of text (prompt plus model-generated conversation history) an LLM can pay attention to at once when generating a response.

Conversational Coherence. How well a dialogue stays on topic, maintains logical connections across turns, and respects prior context in the conversation.

CWI (College of Western Idaho). A public community college in Idaho that offers associate degrees, certificates, and workforce training.

DDR (Design and Development Research). A systematic research approach focused on studying the processes and outcomes of designing, developing, and evaluating instructional products, tools, or models.

Design Fidelity. The degree to which a prototype or implementation matches the intended final design in features, look-and-feel, and functionality.

Design Principle. A generalizable rule or guideline distilled from design and evaluation work that can inform future designs in similar contexts.

Diagnostic Questioning. Asking targeted questions designed to reveal a learner's thinking, misconceptions, or specific skill gaps, not just whether an answer is right or wrong.

Directed Content Analysis. A qualitative analysis approach that starts from existing theory or prior research to create initial (a priori) codes, then applies and refines them as you code data.

Error Classification and Remediation. The process of categorizing learner errors into types (for example, conceptual vs. procedural) and providing tailored feedback or activities to address each type.

Evidence-Informed Design. An approach to designing instruction or tools that intentionally draws on empirical research, data from evaluation, and established theory rather than intuition alone.

Expert Review. A type of formative evaluation where subject-matter or design experts systematically review a product or prototype and provide feedback against criteria or standards.

Formative Evaluation. Evaluation conducted during design and development to gather feedback and data that improve the product or intervention before final implementation.

GPT (Generative Pre-trained Transformer). A family of large neural network models trained on massive text corpora to predict the next token, which can then generate coherent text, code, and other outputs from prompts.

Hallucination (AI). When an AI model generates confident-sounding but factually incorrect, fabricated, or logically unsupported content.

Inductive Code. A code created while analyzing the data, when you notice patterns or ideas that were not pre-specified.

Informed Consent. A process where participants are given clear information about a study's purpose, procedures, risks, benefits, and rights and voluntarily agree to take part.

Intelligent Tutoring System (ITS). A computer-based tutoring system that adapts instruction to the learner by tracking performance, modeling knowledge, and giving individualized feedback and guidance.

IRB (Institutional Review Board). A committee that reviews research involving human participants to ensure ethical conduct and protection of participants' rights and welfare.

ISU (Idaho State University). A public research university in Idaho that offers undergraduate and graduate programs, including in education and instructional technology.

Iterative Design. A design approach where you create a version, test or evaluate it, then refine and repeat in cycles to progressively improve the product.

Likert Scale. A rating scale (for example, strongly disagree to strongly agree) used to measure attitudes or perceptions across ordered response options.

LLM (Large Language Model). A very large neural network trained on text to model language; it can generate and transform text, follow instructions, and answer questions in natural language.

Mastery Check Gating. A design where progression to the next unit or activity is blocked until the learner passes a mastery check, such as a quiz or performance criterion.

Mastery Learning. An approach where students are expected to reach a defined level of understanding on each unit, with support and remediation, before moving on.

Math Tutor GPT. A GPT-based system configured to act as a math tutor, providing step-by-step guidance, hints, feedback, and explanations for math problems within defined constraints and prompts.

Motivational Framing. How learning tasks and messages are presented to influence students' interest, value beliefs, and sense of competence or relevance.

Open-Ended Item. An assessment question that allows learners to generate their own response (explanations, steps, reasoning) rather than choosing from fixed options.

Pedagogical. Relating to teaching methods, instructional strategies, and how learning experiences are designed and enacted.

Phase (DDR). A distinct stage in a design and development research project, such as analysis, design, development, implementation, or evaluation, often iterated.

Placement Test. An assessment used to determine a student's current skill level so they can be placed into an appropriate course, such as developmental vs. college-level math.

Prompt Engineering. The practice of designing and refining instructions, examples, and formatting in prompts to reliably elicit desired outputs and behaviors from an LLM.

Protocol. A predefined and documented set of procedures, steps, or scripts that guide how a study, observation, or tutoring interaction is carried out.

Purposive Sampling. A non-random sampling strategy where you deliberately select participants or cases because they are especially informative for your research questions.

Qualtrics. An online survey and experience-management platform used to design, distribute, and analyze surveys and related data.

Remedial Mathematics. Math courses or supports designed to help students who are not yet ready for credit-bearing college math build foundational skills.

Remediation. Instructional interventions aimed at addressing learning gaps or misconceptions so that students can meet target competencies.

Retrieval Practice. Learning activities that require learners to actively recall information from memory, which strengthens long-term retention more than rereading.

Richey and Klein DDR Framework. A seminal framework that defines DDR as the systematic study of design, development, and evaluation processes and distinguishes between product/tool research and model research.

Scaffolded Hint Ladder. A structured sequence of hints that become progressively more explicit, helping learners move from a brief nudge to a near-complete solution as needed.

Scaffolding. Instructional support that helps learners perform tasks they cannot yet do alone, which is gradually removed as competence increases.

SLTDF (Scaffolded LLM Tutoring Design Framework). A design framework that organizes how to use LLMs for tutoring with purposeful scaffolds, prompting structures, and guardrails aligned to learning theory and DDR-style evaluation.

Spaced Retrieval Prompts. Retrieval practice activities scheduled over time with increasing intervals to leverage the spacing effect and improve durable learning.

System Prompt. The hidden or high-level instruction text given to an LLM that defines its role, goals, constraints, and general behavior before user messages are processed.

Tutor Behavior Specification. A written description of how an AI tutor should behave: what it should say, how it should scaffold, what it must avoid, and how it handles errors, used to guide prompts and design.

Type 1 Study (DDR). A design and development research study focused on specific product or program development and its iterative design, development, and evaluation.

Type 2 Study (DDR). A design and development research study focused on developing or refining broader design or development models, principles, or frameworks.

Worked Examples with Fading. An instructional sequence that starts with fully worked solutions, then gradually removes steps so learners take over more of the problem-solving process.

Zone of Proximal Development. The range of tasks a learner cannot do independently but can accomplish with guidance from a more knowledgeable other (or tutor).

Appendix D: Tutor Behavior Specification

Math Tutor GPT

Appendix D: Tutor Behavior Specification

Andrew Egbert | Idaho State University — EDLT 7733 | Spring 2026

Purpose. This document maps each of the seven evidence-informed tutoring principles instantiated in the Math Tutor GPT to its (1) theoretical basis, (2) observable behavior specification, (3) triggering condition, and (4) expert reviewer validation criterion. It serves simultaneously as the design specification, development guide, and primary evaluation instrument backbone for Phase 3 expert review. Expert reviewers should use the validation criteria to assess behavioral presence and appropriateness in each corresponding evaluation scenario.

Use. Each numbered feature below corresponds to one scenario in the Expert Review Instrument (Appendix E). Reviewers should read the relevant feature block before engaging with the corresponding scenario.

1. Diagnostic Questioning

Theoretical Basis	• Ausubel (1968) — prior knowledge activation • Corbett & Anderson (1994) — knowledge tracing • Shute & Towle (2003) — adaptive instruction
Observable Tutor Behavior	At session initiation, tutor presents 3–5 targeted items spanning arithmetic, pre-algebra, and algebra. Learner responses route them to an appropriate starting point in the content sequence. Tutor explicitly acknowledges diagnostic purpose and references diagnostic results when selecting content.
Triggering Condition	First learner message of a session. OR learner explicitly requests to begin a new topic or start over.
Expert Reviewer Validation Criteria	✓ Diagnostic items span at least two skill levels ✓ Tutor explicitly links diagnostic responses to routing decision ✓ Content delivery does not begin before diagnostic is conducted ✓ Tutor references prior knowledge profile when explaining sequencing choices

2. Scaffolded Hint Ladder

Theoretical Basis	• Vygotsky (1978) — Zone of Proximal Development • Belland (2017) — scaffolding fading • VanLehn (2006) — ITS hint architecture
Observable Tutor Behavior	On error or hint request, tutor responds at the conceptual tier first (e.g., “What operation relates these quantities?”). Second request yields procedural step-level hint. Third request yields a fully worked step. Tiers are not skipped in either direction.
Triggering Condition	Learner submits an incorrect answer. OR learner explicitly requests a hint.

Expert Reviewer Validation Criteria	✓ First hint is conceptual, not procedural or worked ✓ Subsequent hints escalate in specificity across ≥2 tiers ✓ Fully worked step is withheld until prior tiers are exhausted ✓ Individual tiers are content-distinct (not paraphrases of each other)

3. Worked Examples with Fading

Theoretical Basis	• Sweller & Cooper (1985) — worked examples effect • Renkl & Atkinson (2003) — example-fading sequence • Sweller et al. (1998) — Cognitive Load Theory
Observable Tutor Behavior	For a new problem type: tutor presents a fully annotated worked example first. Then presents a partially completed example requiring learner input on 1–2 steps. Then requires fully independent problem solving. Fading sequence is not reversed.
Triggering Condition	Learner encounters a new problem type or topic for the first time in the session. OR learner states unfamiliarity with a concept. OR diagnostic results indicate no prior exposure.
Expert Reviewer Validation Criteria	✓ Full worked example appears before any independent practice ✓ Partial example explicitly requires learner to complete missing steps ✓ Annotation/explanation accompanies worked steps (not just notation) ✓ Tutor does not present full example and immediately require full independent performance

4. Mastery Check Gating

Theoretical Basis	• Bloom (1968) — mastery learning criterion • Guskey (2007) — criterion-referenced progression • Corbett & Anderson (1994) — mastery threshold in ITS
Observable Tutor Behavior	Before introducing a new topic or sub-skill, tutor requires learner to correctly solve ≥2 similar problems without scaffolded support. Learner failing criterion receives additional targeted practice. Tutor does not advance until criterion is met. Gating logic is made explicit to the learner.
Triggering Condition	Learner appears to have completed structured practice on a topic area. Tutor assesses readiness before introducing new skill domain.

Expert Reviewer Validation Criteria	✓ Topic advancement does not occur without demonstrated mastery ✓ Mastery criterion requires ≥2 correct unscaffolded responses ✓ Learner who fails criterion receives targeted remediation, not advancement ✓ Tutor communicates the mastery requirement and rationale to the learner

5. Spaced Retrieval Prompts

Theoretical Basis	• Cepeda et al. (2006) — spacing effect • Roediger & Karpicke (2006) — retrieval practice • Dunlosky et al. (2013) — distributed practice
Observable Tutor Behavior	After moving on from a mastered topic to new content, tutor inserts review problems from previously covered domains. Review prompts are labeled as such and explicitly reference prior coverage. Tutor provides rationale (memory consolidation) when introducing review.
Triggering Condition	Learner has demonstrated mastery on a prior topic. AND session has progressed through at least 1 additional topic area. Tutor triggers retrieval probe before beginning a new topic.
Expert Reviewer Validation Criteria	✓ Retrieval prompts occur after topic transition, not only within-topic ✓ Review items explicitly reference prior coverage (“Earlier we worked on…”) ✓ Tutor provides rationale for the review prompt ✓ Review items target previously mastered content at appropriate difficulty

6. Error Classification & Remediation

Theoretical Basis	• Hattie & Timperley (2007) — feedback theory • Anderson et al. (1995) — Cognitive Tutor feedback model • Booth et al. (2013) — error analysis in mathematics
Observable Tutor Behavior	On error, tutor classifies the error as conceptual (misunderstanding of underlying principle), procedural (correct concept but execution error), or careless (likely arithmetic slip). Conceptual errors receive explanatory feedback targeting the underlying principle. Procedural errors receive step-level corrective guidance. Careless errors receive a brief correction prompt. Feedback content differs substantively across the three error classifications.
Triggering Condition	Learner submits an incorrect answer to a practice problem. OR a learner answer is detected as inconsistent with the expected solution.

Expert Reviewer Validation Criteria	✓ Tutor does not issue generic “incorrect, try again” feedback ✓ Tutor characterizes or names the error type before responding ✓ Feedback content is substantively different across error classifications ✓ Conceptual errors receive more elaborated explanation than procedural/careless errors

7. Motivational Framing

Theoretical Basis	• Dweck (2006) — growth mindset • Dweck & Leggett (1988) — implicit theories of intelligence • Hattie & Timperley (2007) — feedback theory (motivation focus)
Observable Tutor Behavior	Tutor frames effort and errors with growth-oriented language. Praises effort and process rather than fixed ability. Normalizes struggle as part of learning rather than evidence of inability. Reframes errors as information about what to practice next, not as failures. Avoids evaluative language that signals permanent inability.
Triggering Condition	Any learner-tutor interaction; motivational framing operates as an ambient stance.
Expert Reviewer Validation Criteria	✓ Tutor uses growth-oriented language and process-focused praise ✓ Tutor normalizes struggle and reframes errors as learning opportunities ✓ Tutor avoids fixed-ability or evaluative language ✓ Praise targets effort and process, not fixed ability

---

Note. *The validation criteria are binary/partial assessments of design fidelity, not measures of learning efficacy. This document does not assume that behavioral fidelity to the specification produces learning gains; that claim requires empirical testing in subsequent dissertation phases.*

Appendix E: Expert Review Instrument

Survey Flow Overview

- Introduction & Instructions (Q1)
- Informed Consent (Q44 consent text, Q45 agree/disagree)
- Section A: Reviewer Profile (Q3–Q8)
- Section B: Scenario-Based Evaluation (Q9–Q37, seven scenarios × four items each)
- Scenario 1: Diagnostic Questioning (Q9, Q11, Q12, Q13)
- Scenario 2: Scaffolded Hint Ladder (Q14, Q15, Q16, Q17)
- Scenario 3: Worked Examples with Fading (Q18, Q19, Q20, Q21)
- Scenario 4: Mastery Check Gating (Q22, Q23, Q24, Q25)
- Scenario 5: Spaced Retrieval Prompts (Q26, Q27, Q28, Q29)
- Scenario 6: Error Classification & Remediation (Q30, Q31, Q32, Q33)
- Scenario 7: Motivational Framing (Q34, Q35, Q36, Q37)
- Section C: Global Behavioral Ratings (Q38, Q39)
- Section D: Holistic Evaluation (Q40 = D1 fidelity, Q41 = D2 strengths, Q42 = D3 weaknesses, Q43 = D4 revisions, Q44 = D5 acknowledgment)

Introduction & Instructions

Study purpose. This instrument supports the formative expert review of the Math Tutor GPT, an LLM-based remedial math tutoring tool designed for students preparing for the College of Western Idaho mathematics placement assessment. The tutor operationalizes seven evidence-informed tutoring behaviors derived from the instructional design and cognitive tutoring literature. The purpose of this expert review is to assess design fidelity of the implemented tutor relative to its behavior specification; that is, to determine whether each intended behavior is present, and whether each implemented behavior is appropriate relative to its theoretical basis.

What you are NOT being asked to evaluate. You are not being asked to assess learning outcomes; this study does not include pre/post learning data, general quality impressions, or commercial potential. You are evaluating the degree to which the tutor exhibits the specific behaviors it was designed to exhibit.

Time required. Approximately 60 to 90 minutes. Seven scenarios are included, each requiring 5 to 10 minutes of interaction with the tutor and 5 minutes of rating time. Global evaluation items require an additional 10 to 15 minutes.

Access. The Math Tutor GPT is accessible via ChatGPT at: https://chatgpt.com/g/g-694dc0a9f0008191b63dc60f687f2ccc-cwi-math-tutor (opens in a new window). You will need a free or paid ChatGPT account. Open a new chat for each scenario, or use the Reset conversation option between scenarios to avoid carryover effects.

Confidentiality. Your responses will be used solely for formative evaluation of this research prototype. Your name or affiliation will not be associated with individual

ratings in any publication. If you consent to being acknowledged in the final report, you may opt in at the end of this instrument.

Behavior Specification Reference. Before beginning each scenario, consult Appendix D: Tutor Behavior Specification. Each scenario maps to one row of that table.

Informed Consent

Study Title: Expert Evaluation of an AI-Powered Math Tutoring System

Principal Investigator: Andrew Egbert, Idaho State University

Contact: andrewegbert@cwi.edu | 832-475-5213

Purpose of the Study

You are being invited to participate in a research study to evaluate the behavioral accuracy and appropriateness of an AI-powered math tutoring system. The purpose of this study is to gather expert feedback on whether the AI tutor's instructional behaviors align with established evidence-based practices in mathematics education.

What You Will Be Asked to Do

If you agree to participate, you will be asked to review a series of instructional scenarios produced by the AI math tutor and rate each scenario using a structured review instrument. You will evaluate whether specific teaching behaviors are present and appropriate, and provide optional written recommendations. This will take approximately 45 to 60 minutes.

Risks and Benefits

There are no anticipated risks beyond those of everyday professional activity. There is no direct benefit to you, though your expert feedback will contribute to improving AI-assisted mathematics instruction and to the body of research on educational technology design.

Confidentiality

Your participation is anonymous. You will not be asked to provide your name unless you choose to do so. No personally identifiable information will be collected or

linked to your responses. Data will be stored securely and accessible only to the research team. Results may be published or presented in aggregate form, with no individual responses identifiable.

Voluntary Participation

Your participation is entirely voluntary. You may decline to answer any question and may withdraw at any time without penalty.

Questions

If you have questions about this study, please contact the principal investigator at andrewegbert@cwi.edu. For questions about your rights as a research participant, contact the Idaho State University IRB at humsubj@isu.edu or 208-282-2179.

By selecting I Agree below, you confirm that you have read and understood the information above, that you are 18 years of age or older, and that you voluntarily agree to participate.

Q45 — Do you agree to participate in this study?

- I Agree — I have read the informed consent information and agree to participate
- I Do Not Agree — I do not wish to participate and will exit the survey

Skip logic: If "I Do Not Agree" is selected, the survey ends.

Section A: Reviewer Profile

Q3 — Professional Role / Title (Optional)

_____

_____

Q4 — Institution / Organization

_____

_____

Q5 — Primary Area of Expertise (select all that apply)

- ○ Instructional Design
- ○ Learning Science
- ○ Mathematics Education
- ○ Remedial Mathematics
- ○ Community College Education
- ○ Student Success
- ○ Artificial Intelligence / LLM Tool Design
- ○ Other (specify below)

_____

_____

Q6 — Years of Relevant Professional Experience

_____

_____

Q7 — Familiarity with LLM-Based Educational Tools

- ○ No prior experience
- ○ Some familiarity (have used, but not designed)
- ○ Moderate (have designed or evaluated LLM tools)
- ○ Extensive (research or development background)

Q8 — Conflict of Interest Disclosure (optional): Any relationship to the researcher or CWI that could bias your evaluation? If none, leave blank.

_____

_____

_____

_____

_____

_____

Section B: Scenario-Based Evaluation

Complete all seven scenarios in order. For each scenario: (1) read the evaluation objective and setup instructions; (2) interact with the Math Tutor GPT following the scripted procedure; (3) complete the rating items immediately after the scenario while the interaction is fresh. Start a new or reset chat session for each scenario.

Scenario 1: Diagnostic Questioning

Q9 — Scenario task description

Evaluation Objective. Determine whether the tutor conducts a structured prior knowledge assessment before initiating content delivery.

Setup. Open a fresh session with the Math Tutor GPT. Do not volunteer any information about your math background.

Procedure

- Type exactly: "Hi, I want to practice for my math placement test."
- Observe whether the tutor initiates a diagnostic protocol or proceeds directly to content.
- If diagnostic items are presented, respond with a mix of correct and incorrect answers across skill levels.
- Note whether the tutor (a) acknowledges your responses diagnostically, and (b) explains why it is starting at a particular content level.

What to observe

- Does the tutor ask about your background before presenting content?
- Does it present items spanning multiple skill levels (not just one)?
- Does it explicitly link your diagnostic responses to its routing decision?

Q11 — S1 Behavioral Presence: Is the Diagnostic Questioning behavior exhibited by the tutor?

- ○ Yes — fully exhibited
- ○ Partial — exhibited but incomplete or inconsistent
- ○ No — not exhibited

Q12 — S1 Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

- ○ 1 — Not appropriate (seriously inconsistent with theoretical basis)

- 2 — Somewhat appropriate (partially consistent; notable gaps)
- 3 — Mostly appropriate (largely consistent; minor gaps)
- 4 — Fully appropriate (strongly consistent; aligns with research base)
- NA — Behavior not present

Q13 — S1 Revision Recommendations: What should change, and why? (Optional)

_____

_____

_____

_____

_____

_____

_____

_____

Scenario 2: Scaffolded Hint Ladder

Q14 — Scenario task description

Evaluation Objective. Determine whether the tutor delivers hints in a tiered sequence (conceptual to procedural to worked step) without skipping tiers.

Setup. After any diagnostic or warm-up, ask the tutor to give you a practice problem in pre-algebra (e.g., solving a one-step equation).

Procedure

- When the problem is presented, submit a clearly incorrect answer (e.g., for 2x = 10, answer 4).
- Type: "Can you give me a hint?" Observe the first hint — note whether it is conceptual or procedural.
- Type: "I still don’t understand, can you give me another hint?" Observe whether the second hint escalates in specificity.
- Type: "I need more help." Observe whether a fully worked step is provided.

What to observe

- Is the first hint conceptual (targeting understanding, not a worked step)?
- Does hint specificity escalate across requests?
- Is a fully worked step withheld until at least two prior hints have been given?

Q15 — S2 Behavioral Presence: Is the Scaffolded Hint Ladder behavior exhibited by the tutor?

- ○ Yes — fully exhibited
- ○ Partial — exhibited but incomplete or inconsistent
- ○ No — not exhibited

Q16 — S2 Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

- ○ 1 — Not appropriate (seriously inconsistent with theoretical basis)

- 2 — Somewhat appropriate (partially consistent; notable gaps)
- 3 — Mostly appropriate (largely consistent; minor gaps)
- 4 — Fully appropriate (strongly consistent; aligns with research base)
- NA — Behavior not present

Q17 — S2 Revision Recommendations: What should change, and why? (Optional)

_____

_____

_____

_____

_____

_____

_____

_____

Scenario 3: Worked Examples with Fading

Q18 — Scenario task description

Evaluation Objective. Determine whether the tutor follows a full to partial to independent example-fading sequence when introducing a new problem type.

Setup. Tell the tutor you want to learn how to solve two-step equations for the first time.

Procedure

- Type: "I have never done two-step equations before. Can you teach me?"
- Observe whether the tutor presents a fully annotated worked example first.
- Note whether the next task is a partial-completion example or an immediate fully independent problem.
- If a partial example is presented, complete it and observe whether full independent practice follows.
- Note whether each transition in the fading sequence is labeled or explained.

What to observe

- Is a fully annotated worked example presented before any independent practice?
- Is there a partial-completion example between the full example and independent practice?
- Does the tutor label or explain the transitions in the fading sequence?

Q19 — S3 Behavioral Presence: Is the Worked Examples with Fading behavior exhibited by the tutor?

- ○ Yes — fully exhibited
- ○ Partial — exhibited but incomplete or inconsistent
- ○ No — not exhibited

Q20 — S3 Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

- 1 — Not appropriate (seriously inconsistent with theoretical basis)
- 2 — Somewhat appropriate (partially consistent; notable gaps)
- 3 — Mostly appropriate (largely consistent; minor gaps)
- 4 — Fully appropriate (strongly consistent; aligns with research base)
- NA — Behavior not present

Q21 — S3 Revision Recommendations: What should change, and why? (Optional)

_____

_____

_____

_____

_____

_____

Scenario 4: Mastery Check Gating

Q22 — Scenario task description

Evaluation Objective. Determine whether the tutor requires mastery demonstration before advancing to a new topic.

Setup. Work through several practice problems on a single topic (e.g., fraction operations). Answer about half correctly and half incorrectly initially, then answer the last two correctly.

Procedure

- Engage with 4 to 6 practice problems on a single topic.
- Observe whether the tutor attempts to advance to a new topic after correct answers.
- Note whether advancement is gated by a mastery criterion or occurs arbitrarily.
- To test gating, deliberately answer incorrectly after an initial success streak. Observe whether the tutor holds position and provides additional practice.

What to observe

- Does the tutor explicitly gate topic advancement on demonstrated mastery?
- Is the criterion stated or implied?
- Does the tutor provide additional practice rather than advancing when mastery is not met?

Q23 — S4 Behavioral Presence: Is the Mastery Check Gating behavior exhibited by the tutor?

- Yes — fully exhibited
- Partial — exhibited but incomplete or inconsistent
- No — not exhibited

Q24 — S4 Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

- 1 — Not appropriate (seriously inconsistent with theoretical basis)
- 2 — Somewhat appropriate (partially consistent; notable gaps)
- 3 — Mostly appropriate (largely consistent; minor gaps)
- 4 — Fully appropriate (strongly consistent; aligns with research base)
- NA — Behavior not present

Q25 — S4 Revision Recommendations: What should change, and why? (Optional)

_____

_____

_____

_____

_____

_____

Scenario 5: Spaced Retrieval Prompts

Q26 — Scenario task description

Evaluation Objective. Determine whether the tutor inserts retrieval practice from previously covered material after progressing to new content.

Setup. Work through two complete topics to mastery (e.g., order of operations, then fraction operations). Do not mention prior topics after transitioning.

Procedure

- Complete the first topic to mastery. Transition to the second topic.
- Complete several problems on the second topic.
- Observe whether the tutor inserts a problem from the first topic mid-session.
- Note whether the tutor labels the prompt as review and provides a rationale.

What to observe

- Does a retrieval prompt from a prior topic appear?
- Is it labeled as review?
- Does the tutor explain why it is returning to prior content?

Q27 — S5 Behavioral Presence: Is the Spaced Retrieval Prompts behavior exhibited by the tutor?

- Yes — fully exhibited
- Partial — exhibited but incomplete or inconsistent
- No — not exhibited

Q28 — S5 Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

- 1 — Not appropriate (seriously inconsistent with theoretical basis)
- 2 — Somewhat appropriate (partially consistent; notable gaps)

- 3 — Mostly appropriate (largely consistent; minor gaps)
- 4 — Fully appropriate (strongly consistent; aligns with research base)
- NA — Behavior not present

Q29 — S5 Revision Recommendations: What should change, and why? (Optional)

_____

_____

_____

_____

_____

_____

_____

_____

Scenario 6: Error Classification & Remediation

Q30 — Scenario task description

Evaluation Objective. Determine whether the tutor classifies student errors by type and provides remediation matched to the identified error type.

Setup. Work through practice problems intentionally making different types of errors (e.g., a procedural error such as wrong operation order; a conceptual error such as misapplying a rule).

Procedure

- Commit a clear procedural error (e.g., add instead of subtract in a two-step equation).
- Observe whether the tutor identifies the type of error made.
- Commit a conceptual error (e.g., treat a fraction incorrectly as a whole number).
- Observe whether the tutor responds differently to conceptual versus procedural errors.

What to observe

- Does the tutor identify the error type?
- Does the remediation strategy match the error type?
- Does the tutor distinguish between procedural and conceptual errors?

Q31 — S6 Behavioral Presence: Is the Error Classification & Remediation behavior exhibited by the tutor?

- Yes — fully exhibited
- Partial — exhibited but incomplete or inconsistent
- No — not exhibited

Q32 — S6 Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

- 1 — Not appropriate (seriously inconsistent with theoretical basis)
- 2 — Somewhat appropriate (partially consistent; notable gaps)
- 3 — Mostly appropriate (largely consistent; minor gaps)
- 4 — Fully appropriate (strongly consistent; aligns with research base)
- NA — Behavior not present

Q33 — S6 Revision Recommendations: What should change, and why? (Optional)

_____

_____

_____

_____

_____

_____

Scenario 7: Motivational Framing

Q34 — Scenario task description

Evaluation Objective. Determine whether the tutor uses effort-based, growth-oriented feedback language rather than ability-based or evaluative language.

Setup. Engage with the tutor across several problems, deliberately answering some correctly and some incorrectly.

Procedure

- Answer a problem correctly. Note the tutor’s feedback.
- Answer a problem incorrectly. Note the tutor’s feedback.
- Express frustration: Type "I feel like I’m just bad at math."
- Observe whether the tutor reframes the statement in effort or growth terms.

What to observe

- Does feedback emphasize effort and strategy rather than innate ability?
- Does the tutor avoid language like "you’re smart" or "that was easy"?
- Does the tutor respond to negative self-talk with growth-oriented reframing?

Q35 — S7 Behavioral Presence: Is the Motivational Framing behavior exhibited by the tutor?

- Yes — fully exhibited
- Partial — exhibited but incomplete or inconsistent
- No — not exhibited

Q36 — S7 Behavioral Appropriateness: Is the exhibited behavior consistent with its theoretical basis? (Skip if Presence = No)

- 1 — Not appropriate (seriously inconsistent with theoretical basis)
- 2 — Somewhat appropriate (partially consistent; notable gaps)

- 3 — Mostly appropriate (largely consistent; minor gaps)
- 4 — Fully appropriate (strongly consistent; aligns with research base)
- NA — Behavior not present

Q37 — S7 Revision Recommendations: What should change, and why? (Optional)

_____

_____

_____

_____

_____

_____

_____

_____

Section C: Global Behavioral Ratings

Q38 — For each of the seven behaviors, rate the overall appropriateness of the tutor's implementation. Base your rating on observations across all interactions, not just the specific scenario task.

Behavior	1 Not appropriate	2 Somewhat	3 Mostly	4 Fully
B1: Diagnostic Questioning	○	○	○	○
B2: Scaffolded Hint Ladder	○	○	○	○
B3: Worked Examples with Fading	○	○	○	○
B4: Mastery Check Gating	○	○	○	○
B5: Spaced Retrieval Prompts	○	○	○	○
B6: Error Classification & Remediation	○	○	○	○
B7: Motivational Framing	○	○	○	○

Q39 — Global Comments: Are there any cross-cutting patterns, design strengths, or concerns that you observed across multiple behaviors? (Optional)

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Section D: Holistic Evaluation

Q40 — D1. Overall Design Fidelity Rating: Considering all seven behaviors in aggregate, how would you rate the overall design fidelity of the Math Tutor GPT relative to its behavior specification?

- 1 — Low fidelity (fewer than 3 behaviors fully or partially exhibited)
- 2 — Partial fidelity (3 to 4 behaviors exhibited; notable gaps remain)
- 3 — Moderate fidelity (5 to 6 behaviors exhibited; minor gaps)
- 4 — High fidelity (all 7 behaviors exhibited and appropriately implemented)

Q41 — D2. Primary Strengths: What design features or implemented behaviors did you find most consistent with the theoretical literature? Be specific.

_____
_____

_____
_____

_____
_____

_____
_____

_____
_____

_____

_____

Q42 — D3. Primary Weaknesses: What design gaps, inconsistencies, or theoretical misalignments were most notable? Be specific.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Q43 — D4. Priority Revision Recommendations: If you could identify three specific design changes that would most improve the tutor's alignment with its theoretical basis, what would they be?

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Q44 — D5. Acknowledgment Consent: May the researcher acknowledge your participation (first name, last name, and institutional affiliation) in the final dissertation report and any resulting publications?

- Yes — you may acknowledge me by name and affiliation
- No — please keep my participation confidential

Appendix F: Reviewer Orientation Webpage

This appendix reproduces the content of the reviewer orientation webpage that was distributed to expert reviewers. The live page is available at https://appliedarts.us.com/mathtutorreview/reviewer_page.html and serves as the asynchronous protocol brief that frames the study, presents the seven design features, outlines the three scripted scenarios, and links to the Math Tutor GPT and the Qualtrics survey.

Header

Expert Review Protocol

CWI Math Tutor GPT: Design and Development Study

A formative expert review of a theory-guided AI tutoring system designed to prepare community college developmental math students for placement assessments.

Researcher. Andrew Egbert, M.S. Doctoral Student, Instructional Design and Technology, ISU.

Institution. College of Western Idaho (CWI).

Study Type. Type 1 Design and Development Research (Richey and Klein, 2007).

Overview: What You Are Being Asked to Review

This study is a formative expert evaluation of the CWI Math Tutor GPT (Version 4.1), a custom GPT built on OpenAI's platform. The tutor is designed to operationalize seven evidence-informed tutoring behaviors, drawn from cognitive tutoring research, self-regulated learning theory, and instructional design literature, into a consistent, controllable AI tutoring experience.

The tutor's target population is developmental math students at the College of Western Idaho who are preparing for unit placement assessments covering whole numbers, integers, fractions, decimals, basic algebra, and introductory geometry.

Research Context. This study is classified as Type 1 Product and Tool Research per Richey and Klein (2007). The goal is design validation evidence: specifically, expert judgment on whether the tutor's implemented behaviors align with the stated theoretical rationale, whether the design is coherent and usable, and where meaningful gaps or improvements exist. This is formative evaluation, not efficacy testing with students.

Estimated total time: 45 to 60 minutes.

Why Expert Review?

Expert review is an established formative evaluation method in design and development research. Reviewers with relevant expertise, in instructional design, developmental mathematics instruction, or AI/LLM applications in education, are better positioned than naive users to evaluate the coherence between the tutor's design rationale and its actual behavior, identify pedagogically significant failure modes, and offer actionable improvement recommendations.

Reviewer feedback directly informs the next iteration of the tutor and contributes to a transferable framework for designing pedagogically grounded LLM-based tutors in community college contexts.

Informed Consent

Voluntary Participation

Participation in this expert review is entirely voluntary. Reviewers may withdraw at any time without consequence. No personally identifying information will be reported; findings are presented in aggregate. This study has been submitted to the Idaho State University Institutional Review Board (IRB) for exempt determination under 45 CFR 46.104(d)(2).

By proceeding with the review and submitting the Qualtrics survey, reviewers acknowledge that they have read this notice, understand the nature of the study, and voluntarily agree to participate.

Questions or Concerns. If reviewers have questions about their rights as a research participant, they may contact the ISU Human Subjects Committee at

humsubj@isu.edu or 208-282-2179. For questions about the study itself, reviewers contact the researcher at the address listed at the bottom of the page.

Seven Evidence-Informed Tutoring Behaviors

The CWI Math Tutor GPT is designed to operationalize the following seven tutoring behaviors, each grounded in specific learning science literature. Understanding these behaviors helps reviewers evaluate whether the tutor's responses reflect its intended design and where implementation falls short.

01. Diagnostic Questioning. Begins sessions with targeted questions to establish prior knowledge and calibrate difficulty before instruction begins.

02. Scaffolded Hint Ladder. Provides hints in graduated steps: smallest useful hint first, escalating only as needed, with full solution as a last resort.

03. Worked Examples with Fading. Introduces skills through complete worked examples, then progressively reduces support as learner fluency develops.

04. Mastery Check Gating. Requires demonstrated competence (two correct in a row with low support) before advancing to the next topic or skill.

05. Spaced Retrieval Prompts. Reintroduces previously mastered skills after a delay to consolidate long-term retention through distributed practice.

06. Self-Explanation Prompting. Prompts learners to articulate their reasoning and strategy in their own words, supporting metacognitive awareness.

07. Error Diagnosis with Differential Response. Classifies errors (mechanical, conceptual, strategic, overload, symbol confusion) and responds differentially based on error type.

What This Means for Your Review. Reviewers do not need to evaluate all seven behaviors equally in every scenario. Some scenarios are designed to surface specific behaviors. The expert review instrument asks reviewers to rate each behavior separately based on overall interaction experience.

How to Conduct the Review

Step 1. Read this page in full before opening the tutor. Understanding the design rationale makes the evaluation more precise and useful.

Step 2. Open the CWI Math Tutor GPT. A free OpenAI/ChatGPT account is required to interact with a custom GPT. Account creation takes about two minutes at chatgpt.com.

Step 3. Work through the three scripted scenarios below, in order. Each scenario is designed to elicit specific tutoring behaviors. Reviewers should spend approximately 10 to 15 minutes per scenario, taking brief notes as they go.

Step 4. Reviewers may deviate from the scripted prompts to probe interesting behavior. The scenarios are starting points, not rigid scripts. Expert reviewers are expected to use professional judgment.

Step 5. Complete the Qualtrics survey immediately after the interaction sessions while observations are fresh. The survey takes approximately 15 to 20 minutes.

Important Note on LaTeX Rendering. The tutor is configured to render all mathematical expressions in LaTeX. In the ChatGPT interface, this renders as formatted math. If reviewers see raw LaTeX code (for example, \frac{3}{4}), this is a rendering environment limitation, not a tutor error. Reviewers may note it if it affects readability but should not score it as a design failure.

Tutor link: https://chatgpt.com/g/g-694dc0a9f0008191b63dc60f687f2ccc-cwi-math-tutor

Review Scenarios

Reviewers use the following scenarios to structure interaction with the tutor. Each scenario targets a different combination of the seven tutoring behaviors. Reviewers start a fresh conversation for each scenario.

Scenario 1: The Struggling Beginner. Fractions and Error Response

Primary focus. Diagnostic questioning, scaffolded hint ladder, error diagnosis with differential response.

Opening prompt. "Diagnose me for fractions."

What to look for. Does the tutor ask a goal-setting question before beginning? Does it start with an appropriate diagnostic problem rather than jumping into instruction? When the reviewer answers incorrectly (for example, by adding numerators and denominators separately, 1/2 + 1/3 = 2/5), observe how the tutor classifies and responds to the error. Does it give the smallest useful hint first, or does it over-explain immediately?

Behavior tags. Diagnostic Questioning, Hint Ladder, Error Diagnosis

Scenario 2: The Shortcut-Seeker. Mastery Gating and Self-Explanation

Primary focus. Mastery check gating, self-explanation prompting, worked examples with fading.

Opening prompt. "I need to review two-step equations. Just give me the answer steps, I know how to do this."

What to look for. Does the tutor resist simply providing solutions on demand? Does it require an attempt before providing explanation? After two correct responses in a row, does it advance, check for transfer, or prompt the reviewer to explain reasoning? When the reviewer says "just tell me the answer," does the tutor maintain its instructional stance or capitulate?

Behavior tags. Mastery Gating, Self-Explanation, Worked Examples

Scenario 3: The Placement Test Sprint. Mixed Review and Spaced Retrieval

Primary focus. Spaced retrieval prompts, placement-test alignment, session-level coherence.

Opening prompt. "Give me a Unit 0 timed mini-check. I want to see where I stand."

What to look for. Does the tutor select a representative sample of Unit 0 topics rather than clustering on one skill? Does it score by topic, not just total correct? After the mini-check, does it provide a meaningful diagnostic summary and recommend next steps? If the reviewer continues the session for 10 or more minutes, do previously covered topics re-appear in a spaced retrieval context?

Behavior tags. Spaced Retrieval, Placement Alignment, Session Coherence

Reviewer Discretion. After completing all three scenarios, reviewers may continue interacting with the tutor freely to probe any behavior of interest. Additional observations from open-ended interaction are welcome in the survey’s open-response items.

Complete the Expert Review Survey

After completing the interaction sessions with the tutor, reviewers complete the structured review instrument. The survey takes approximately 15 to 20 minutes and covers:

Ratings of each of the seven tutoring behaviors (alignment, consistency, quality)
Overall usability and coherence of the tutoring experience
Specific observations from each scenario
Open-ended recommendations for improvement
Brief background questions (area of expertise, years of experience)

Please complete the survey promptly after the interaction. Observations are most detailed and accurate immediately after the session. Reviewers are advised to open the survey in a second browser tab before beginning the tutor interaction so it is ready when they finish.

Survey link: https://isu.co1.qualtrics.com/jfe/form/SV_3gwcnZsgBBJ28l0

Contact the Researcher

Reviewers with questions about the study design, the tutor's intended behavior, or the review process before or after their session may reach out directly:

Andrew Egbert

Doctoral Student, Instructional Design and Technology

Idaho State University, College of Education

Instructional Designer, College of Western Idaho

andrew.egbert@isu.edu

For questions about a participant's rights, the ISU Human Subjects Committee may be contacted at humsubj@isu.edu or 208-282-2179.