

**Theory-Guided GPT Tutoring and Math Placement Preparation in Community College:
A Quantitative Research Proposal**

Andrew Egbert

Instructional Design and Development, Idaho State University

EDLP 7706: Advanced Research Design II: Quantitative Methods

Dr. Rick Wagoner

May 6, 2026

Theory-Guided GPT Tutoring and Math Placement Preparation in Community College: A Quantitative Research Proposal

Students entering the College of Western Idaho (CWI) are required to engage in math placement testing in CWI's Math Solutions Center to ensure that they are properly aligned with the correct math courses at the beginning of the semester. The assessment procedures often discover that learners are underprepared for the placement tests. Suboptimal results typically route the student into a developmental, remedial course sequence below the typical college-mathematics level. The chance of attrition is increased with each additional remedial course, which often disproportionately affects low-income or first-generation learners (Ngo & Melguizo, 2016).

Preparation in advance of placement testing can be the key to success, helping students to boost their scores enough to avoid additional remedial courses. AI Large Language Models (LLMs) such as ChatGPT and Claude present low-cost remediation opportunities that offer on-demand, individualized instructional dialogue (Mollick & Mollick, 2023). ChatGPT successfully demonstrated its capacity to generate learning outcomes that rival or improve upon human tutoring in mathematics topics (Pardos & Bhandari, 2024). However, this does not guarantee instructionally effective tutoring, and VanLehn (2011) argued that Intelligent Tutoring Systems (ITS) are made more effective when they are designed around scaffolding and step-based feedback rather than raw computational power. Additionally, the Interactive-Constructive-Active-Passive (ICAP) framework developed by Chi and Wylie (2014) suggests that constructive and interactive engagement, rather than passive intake of correct information, increases learning.

Research and empirical evidence comparing theory-guided LLM tutors to unstructured LLM tutors in community college placement testing is extremely limited. That is, previous investigations into LLM-based learning-gain work do not address these theoretical mechanisms (Pardos & Bhandari, 2024), while preferring to measure intent-to-use or learner perceptions rather than outcomes derived from high-stakes assessments (Mollick & Mollick, 2023). Therefore, findings derived from the comparison of theory-guided Generative Pre-trained Transformer (GPT) AI tutors with baseline GPT tutors, in conjunction with an interaction-level regression model, would contribute by delivering data that helps to answer which design features matter (the instructional design question) and which mechanisms produce the greatest effect (the mechanism question).

Purpose of the Study

This study has dual components. First, the study will compare posttest performance between community college students who use a theory-guided GPT-based math tutor and those who use a baseline GPT-based math tutor to prepare for placement testing. Second, the study will determine which theory-guided GPT-based math-tutor variables predict posttest scores, while controlling for pretest performance. Variables include the frequency of scaffolded hints, the mastery-check completion rate, and the number of attempts before the solution is revealed.

Significance of the Study

This research directly addresses an instructional design question with direct implications for community college access. Should a theory-guided GPT tutor produce better results than a non-theory-guided tutor in a genuine preparation setting, the findings would support theory-guided GPT product design principles and investment in AI-

mediated preparation products, distinguishing such tools from general-purpose or default chatbots.

Interaction variables will be modeled as predictors while contributing to the literature that supports observable learner behaviors with cognitive engagement (Chi & Wylie, 2014) and step-based tutoring (VanLehn, 2011). The study will collect evidence about which pedagogical features most effectively support placement-aligned improvements in a community college context.

Research Questions and Hypotheses

This study asks two research questions.

Research Question 1 (RQ1):

Do students who use a theory-guided GPT-based math tutor demonstrate significantly higher placement-aligned posttest math scores than students who use a baseline GPT tutor, controlling for pretest performance?

Hypothesis 1 (H1):

Students in the theory-guided tutor condition will score significantly higher on the placement-aligned posttest than students in the baseline tutor condition, controlling for pretest performance.

Research Question 2 (RQ2):

To what extent do learner-tutor interaction variables, specifically scaffolded hint-use frequency, mastery-check completion rate, and number of attempts before solution reveal, predict placement-aligned posttest math scores, controlling for pretest performance, among students assigned to the theory-guided tutor condition?

Hypothesis 2a (H2a):

Scaffolded hint-use frequency will positively predict placement-aligned posttest scores.

Hypothesis 2b (H2b):

Mastery-check completion rate will positively predict placement-aligned posttest scores.

Hypothesis 2c (H2c):

The number of attempts before the solution is revealed will positively predict placement-aligned posttest scores, reflecting deeper effortful practice.

Review of the Literature

The literature is focused on three areas, connected by a theoretical framework. First, GPT-based tutoring in mathematics, second, scaffolding and cognitive load, and third, tutoring effectiveness and cognitive engagement.

GPT-Based Tutoring in Mathematics

Empirical evidence for LLM-based mathematics tutors is expanding despite a modest body of literature. In a 274-participant study, Pardos and Bhandari (2024) tested a random comparison of GPT-based hints versus hints created by human tutors. Learning gains were statistically equivalent across the two conditions on elementary and intermediate algebra and statistics. Initial hint quality failures (32% across all topics) were reduced to near zero on algebra problems and 13% on statistics problems through a self-consistency mitigation procedure. Quality control mechanisms were shown to provide a baseline for AI-generated, evidence-based mathematical assistance. Both are woven into the design of this proposal.

Another approach, the Tutor CoPilot study, was carried out by Wang et al. (2024), in which they paired human tutors with an LLM-supported layer in real-time mathematics instruction. The intervention is credited with modest improvements in tutor-delivered learning outcomes. Additionally, it narrowed the gap in effects between highly and lower-rated tutors, showing that LLM-mediated tutoring is most effective when instructional quality is otherwise uneven. The current study attempts to describe a parallel inference. That is, a theory-informed GPT tutor should be most beneficial in placement preparation areas where individualized expert human tutoring support is unavailable to learners.

Evidence has rarely isolated the AI session interaction behaviors that lead to perceived learning gains. Studies have often examined intent-to-use, perceived helpfulness, and student satisfaction rather than learning outcomes tied to high-stakes testing. The proposed study would advance the literature into the community college mathematics placement milieu and has gained scant attention. This study would add an interaction-level mechanism analysis complementary to the outcome comparison in RQ1.

Scaffolding and Cognitive Load Theory

The study's design features provide the conceptual foundation for the theory-guided GPT tutor. Reconceptualizing scaffolding as a distributed system of support across teachers, peer interactions, and tools, Puntambekar (2022) argued that learning is not perpetuated by a single tutor, but by interwoven layers of assistance evolving. This responsibility is gradually transferred from the support system to the learner. However, the tutor must offer sufficient support to keep the learner engaged, though that support should be limited enough to allow the learner to engage personally in the cognitive work. Therefore, the learner is still responsible for their participation and engagement with the

GPT tutor. Puntambekar (2022) highlights the core scaffolding design features, including guidance, diagnosis, fading, and the gradual transfer of responsibility to the learner.

These elements directly translate into the design choices the theory-guided GPT tutor exhibits, including step-based scaffolding, tiered hints that progress from the conceptual to the procedural, mastery checks that assess readiness for advancing to a more advanced topic, and self-explanation prompts that support constructive reasoning.

When considering how scaffolding augments learning rather than simply slowing it down, Sweller (1988) Cognitive Load Theory (CLT) indicates that traditional problem-solving consumes working-memory resources during a means-ends search. It therefore fails to support the acquisition of the schema that enables learners to convert isolated procedures into transferable knowledge structures. In subsequent CLT investigations, van Merriënboer and Sweller (2005) demonstrated that instructional design decisions that reduce extraneous cognitive load while managing intrinsic complexity permit learners to use working-memory capacity for schema construction. This dynamic is moderated by both the learner's motivational state and prior expertise. In short, step-based scaffolding reduces the extraneous load presented by whole-problem solving, tiered hints calibrate intrinsic complexity to learner readiness, and mastery checks generate the diagnostic information that supports the gradual transfer of responsibility. Unlike the baseline condition, theory-guided considerations in the study enable design decisions at the interaction level. The baseline offers unstructured, on-demand conversational assistance, with the cognitive burden of self-pacing and self-diagnosis entirely on the learner.

Tutoring Effectiveness and Cognitive Engagement

Research Question 2 (RQ2) asks whether observable learner-tutor interactions and behaviors carry information about cognitive engagement, and, through that, about learning outcomes. Chi and Wylie (2014)'s Interactive-Constructive-Active-Passive (ICAP) framework suggests that passive engagement is outperformed by a combination of interactive engagement (co-constructive dialog between learners and partners) that evokes deeper learning than constructive engagement, where learners create output that is beyond the given information, which then outperforms active engagement, where learners select or manipulate given information, which then outperforms passive learning. The ICAP framework seeks observable indicators of the foundational engagement mode. ICAP predicts that more constructive and interactive behaviors should elicit stronger learning outcomes.

Bernacki and Walkington (2018) support their findings with empirical anchors, offering their proposition in a math instructional context. They studied an Intelligent Tutoring System (ITS) used for algebra instruction with a sample of 150 randomly selected high school students. They reported that individualized problem contexts triggered greater interest in the learning material. In turn, this triggered further interest in both within-tutor learning efficiency and accuracy, as well as in improved exam outcomes. Bernacki and Walkington (2018)'s found that an observable, measurable feature of the tutor-learner interaction elicits useful predictive information about both immediate and downstream learning outcomes. Additionally, this lends direct empirical support for predictions in this study's RQ2 model.

VanLehn (2011) systematically compared human tutoring, intelligent tutoring systems, and other forms of tutoring and reported that a well-constructed ITS produces

improvements in learning approaching those of one-on-one human tutoring ($d \approx 0.76$ vs. $d \approx 0.79$ for adult human tutors). Additionally, step-level feedback and interactive dialog were found to be key mechanisms that bring ITS to a level approaching human tutoring. This study draws on these findings by operationalizing the frequency of hints and a mastery-check completion scheme, and by using before-solution reveals. These constitute observable proxies for engagement patterns that theory predicts should relate to learning outcomes with a theory-guided GPT tutor.

Theoretical Framework

This study draws on the integration of three theoretical ideas. VanLehn (2011)'s tutoring effectiveness tradition synthesis, Chi and Wylie (2014)'s engagement-as-mechanism ICAP framework, the empirical evidence supplied by Bernacki and Walkington (2018), and cognitive load concepts that originated with Sweller (1988) and enhanced by van Merriënboer and Sweller (2005).

RQ1 is addressed with VanLehn (2011)'s empirical foundation allowing for the expectation of measurable differences between tutoring conditions with differing step-level feedback and scaffolding density. RQ2 is anchored by ICAP's justification that the decision to treat specific learner behaviors (hint requests, mastery-check completion, attempts before solution reveal) is theoretically significant as indicators of cognitive engagement, and not simply usage counts.

Cognitive Load Theory connects both research questions through the judicious application of scaffolding, mastery checks, and tiered hints. Each reduces cognitive load while freeing working memory for schema acquisition. RQ1's design features that are driven by between-condition differences are addressed by the within-condition patterns

addressed by RQ2. Thus, instruction theory is applied to the independent variables as a fundamental theory rather than a post-hoc correlational exploration.

Methods

The study compares a theory-informed GPT tutor with a baseline GPT tutor with a placement-aligned posttest administered after a defined preparation period at the CWI Math Solutions Center. To establish a performance baseline, learners complete a placement-aligned pretest. Once completed, students are granted access to the GPT tutor and then complete a parallel posttest at the conclusion of the exercise.

The tutor automatically generates behavioral log data, which the researcher can unlock using a secret key phrase and copy or download. The data includes event-level records of hint requests, mastery-check item completion, and solution-attempt tallies. If possible, conditions will be randomly assigned, though full randomization may be unfeasible due to the Math Solution Center's operating environment. However, quasi-random assignment within intake time blocks will be used, with pretest scores as a covariate to address non-equivalent baselines.

Population and Anticipated Sample

Incoming CWI students must engage in math placement testing at CWI's Math Solutions Center, and students retaking placement tests are included in this number. These students will be offered the option to participate in the study, which is entirely voluntary, during the research window. Participation is based on admission status, with eligibility restricted to those who are enrolled or admitted to CWI. The study will adhere to the highest ethical considerations, proceeding only after the thorough review and approval by CWI's and ISU's IRBs, as required. Informed consent will include sharing all relevant information with the participant, their explicit agreement to use the GPT tutor

as a study aid, and their consent to the logging and collection of their interaction data. Students may decline participation and will not be compelled. Those who decline will retain full access to all standard Math Solutions Center resources.

The study seeks a sample of 60 to 100 student participants, split as evenly as possible between the two conditions. This will allow the researcher to use a one-way ANOVA to detect a medium effect at the conventional alpha and power levels, and to support multiple regression analysis using a modest predictor set.

Measures and Variables

Measures align with the research questions and are supported by the theoretical framework. Table one defines operational definitions and analytical roles.

Table 1

Study Variables, Operational Definitions, and Analytical Roles

Variable	Operational Definition	Role
Placement-aligned pretest score	Continuous (0–100% correct); items aligned with the CWI placement blueprint, parallel to the posttest	Baseline covariate in RQ1; control variable (Block 1) in RQ2
Placement-aligned posttest score	Continuous (0–100% correct); parallel-form items aligned with CWI blueprint, administered after preparation window	Dependent variable in RQ1 and RQ2
Tutor condition	Dummy coded: 0 = baseline GPT tutor; 1 = theory-guided GPT tutor	Independent variable in RQ1; optional dummy predictor in RQ2 robustness check
Scaffolded hint-use frequency	Count of scaffolded hints requested per student, optionally normalized per item.	Focal predictor in RQ2 (Block 2)
Mastery-check completion rate	Proportion (0–1) of available mastery-check items completed	Focal predictor in RQ2 (Block 2)
Attempts before the	Continuous average of attempts per item	Focal predictor in RQ2 (Block 2)

solution reveal	before the solution is revealed	
Optional controls	Age, first-generation student status, prior math course completion, self-reported test anxiety	Exploratory controls if sample size permits

Note. RQ1 = Research Question 1; RQ2 = Research Question 2.

Instruments

Both pretest and posttest are parallel-form placement-aligned assessments built from current CWI placement tests. Parallel-form construction preserves test coverage, avoiding identical-item exposure across administrations and protecting against memorization that could inflate observed improvements. Tutor interaction logs are produced automatically by both GPT tutors and include real-time stamped records of hint requests, mastery-check item attempts, and solution-reveal requests, which are aggregated into the variables in Table 1.

Analytical Procedures

RQ1 and RQ2 are sequential. Questions and analytical procedures for each are specified in advance. This includes the pre-registered decision rules for RQ1's analysis to avoid post hoc model selection.

Research Question 1: Between-Group Comparison

Means, standard deviations, and ranges will be used to calculate pretest and posttest scores in each condition. This allows the researcher to assess baseline equivalence and determine how outcomes are distributed. Histograms will be examined to assess skewness and kurtosis for normality. Posttest scores will be processed using Levine's test, which will be applied to groups to identify variance, which can be used for ANCOVA. Partial ETA squared (η^2_p) is used for the effect size index and has a 95%

confidence interval. Cohen's d will be used for ANOVA. Effect-size magnitudes are compared with benchmarks from the AI tutoring literature, such as those reported by VanLehn (2011).

Research Question 2: Interaction-Variable Regression

The regression model will be assigned to the theory-guided condition (the production model of Math Tutor GPT), which is reserved for students to avoid mixing interaction behaviors across tutors with divergent coding.

Variables:

- The **dependent variable** is the continuous posttest placement-aligned score.
- The **independent variable**, or Tutor condition (Table 1, Row 3), is the dummy-coded: 0=baseline GPT Tutor; 1=theory-guided GPT tutor.
- The **control variable** is the placement-aligned pretest score (Table 1, Row 1).

The model is hierarchical. Block 1 enters the pretest score as prior-knowledge control. Block 2 adds the three focal interaction predictors: hint-use frequency, mastery-check completion rate, and attempts before solution reveal. Block 3 tests the hint-by-mastery-check interaction.

$$\hat{Y}_0 = \beta_0 + \beta_1(\text{Pretest}) + \beta_2(\text{Hints}) + \beta_3(\text{MasteryRate}) + \beta_4(\text{Attempts}) + \varepsilon$$

Ordinary least squares (OLS) multiple regression will be used. Assumptions will be assessed using scatterplots of each predictor against the outcome for linearity, a conceptual review of the data collection structure (independence of errors), residual plots

(homoscedasticity), normal probability plots (residual normality), and variance inflation factors (multicollinearity), with VIFs above 10 flagged for further examination.

Unstandardized regression coefficients (B), standard errors (SE), and standardized beta weights (β) will be reported for each predictor. Overall model fit will be described using R^2 , and the incremental variance explained by Block 2 relative to Block 1 will be quantified using ΔR^2 . Findings will be situated within the ICAP framework (Chi & Wylie, 2014) and VanLehn's (2011) account of step-based tutoring.

Two supplementary analyses will support robustness. First, a parallel regression including students from both conditions, with a dummy variable for the tutor condition, will assess whether the condition moderates the observed predictor-outcome relationships. Second, a reduced model retaining only the strongest Block 2 predictor will evaluate the stability of the primary findings against predictor-set reduction.

Limitations and Mitigations

Interpretation has three mitigations:

- (1) Recruitment from a single community college limits generalizability.
- (2) Voluntary participation introduces self-selection.
- (3) The small sample size constrains the number of predictors that fit the RQ2 regression, limiting the detection of small effects.

These limitations are addressed through pre-registered decision rules for RQ1, transparent assumption checks, and effect sizes with 95% confidence intervals.

References

- Bernacki, M. L., & Walkington, C. (2018). The role of situational interest in personalized learning. *Journal of Educational Psychology, 110*(6), 864–881.
<https://doi.org/10.1037/edu0000250>
- Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243.
<https://doi.org/10.1080/00461520.2014.965823>
- Mollick, E., & Mollick, L. (2023). *Assigning AI: Seven approaches for students, with prompts*. arXiv. <https://arxiv.org/abs/2306.10052>
- Ngo, F., & Melguizo, T. (2016). How can placement policy improve math remediation outcomes? Evidence from experimentation in community colleges. *Educational Evaluation and Policy Analysis, 38*(1), 171–196.
- Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to those from human-tutor-authored help in mathematics. *PloS one, 19*(5), 1–18. <https://doi.org/10.1371/journal.pone.0304013>
- Puntambekar, S. (2022). Distributed scaffolding: Scaffolding students in classroom environments. *Educational Psychology Review, 34*(1), 451–472.
<https://doi.org/10.1007/s10648-021-09636-3>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147–177.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221.

<https://doi.org/10.1080/00461520.2011.611369>

Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024). *Tutor CoPilot: A Human-AI approach for scaling real-time expertise*. <http://arxiv.org/abs/2410.03017>