RESEARCH REPORT

How the AI Tutor Drives Student Success

Macmillan Learning

AI Tutor

Key Findings:

Our research indicates that the AI Tutor is most effective when students are interactively corresponding with the Tutor through asking questions, responding to pre-entered prompts and working through challenges collaboratively. Learning outcomes peaked when students used the Tutor for about 25% and 75% of their homework questions. In contrast, students who relied too heavily on the Tutor, or limited themselves to only engaging with the pre-entered prompts, realized little to no academic benefits.

Key findings from the research study included:

Improved course performance: Students using the AI Tutor experienced improved assignment, exam and course grades.

Interactive use makes the difference: Students who interactively engaged with the Tutor, rather than using it passively, experienced (statistically and practically) significant learning gains.

Pre-entered prompts were less effective: Students who wrote their own questions and responses to the AI Tutor (instead of relying only on the pre-entered prompts) experienced the greatest gains in understanding and performance. By contrast, those who relied solely on pre-written prompts or clicked through the options without typing messages showed little improvement, and in some cases underperformed compared to peers who didn't use the Tutor at all.

There is an optimal usage pattern: Students who used the Tutor for 25% to 75% of their questions achieved the greatest benefits. Just as minimal or no use offered little advantage, overuse on every question also failed to correlate with better results.

Most students don't overuse: Only 8% of students relied on the Tutor excessively. The research team is continuing to explore this pattern to better understand whether it reflects simple answer-checking behavior, lower confidence, or other underlying factors.

Students used the Tutor, but more students could benefit: Nearly 7 in 10 students engaged with the AI Tutor at least once. Of these, 40% used the tutor regularly across the semester, in a manner that elicited measurable learning gains.

Students value the Tutor's constant availability: The Tutor provides personalized support to all learners, with particular benefits for students who may be reluctant to speak up in class or seek help directedly from their instructor.

Instructors can focus on higher-order learning: Instructors noticed that students were less likely to approach them with basic recall or comprehension questions easily addressed by the AI Tutor. Instead, students sought their instructor's guidance on deeper, more complex problems.

BACKGROUND

Personalized formative feedback is essential to ensuring an equitable learning environment. It fosters self-regulated learning by helping students identify gaps in their knowledge (what they know vs. what they are expected to know) and by guiding them to determine the actions needed to address those gaps (Kluger & DeNisi, 1996; Shute, 2008). Effective formative feedback also delivers high-quality information about students' learning and provides opportunities for students to practice and improve their performance (Kluger & DeNisi, 1996; Shute, 2008).

Student-instructor relationships remain critical for many aspects of learning, including motivation and sense of belonging (Hurtado & Carter, 1997; Ryan & Deci, 2000). However, AI tutors can serve as a valuable supplement by extending the benefits of one-on-one dialogue to students who may have limited access due to time constraints, large class sizes, or class modality constraints. Rather than replacing instructors, generative AI tutoring tools can replicate elements of instructor–student dialogue when appropriate. By drawing on subject matter expertise and effective teaching strategies, the tools are able to provide elaborated feedback and question prompts that promote critical thinking and self-reflection rather than simply providing answers (Shute, 2008).

A growing body of evidence indicates that offering scaffolded, adaptive or AI-based tutoring with personalized feedback enhances engagement and learning in online environments (Doo et al., 2020; Lim et al., 2023). Building on this evidence and recent advances in generative AI capabilities, Macmillan Learning developed and integrated an AI Tutor within Achieve homework assignments. The Tutor is designed to provide personalized, scaffolded feedback to students who may not otherwise have access to high-quality academic support during their learning process.

To determine if the AI Tutor impacted student learning and course performance as intended, Macmillan's department of Learning Science and Research conducted an extensive IRB approved impact research study. Conducting a study across several institutions, with IRB approval, signifies that the study design and methods were independently reviewed and unanimously deemed ethical, credible and valuable to the education and research community. This integrity based certification can not be fabricated or purchased. It can only be obtained by completing the review process.

This research effort spanned the fall 2024 and pring 2025 semesters, across 36 different institutions with 41 instructors, teaching 72 courses. The purpose of the study was to examine whether students who regularly interact with the AI Tutor (within Achieve homework assignments) experience an increase in assignment grade, course exam scores, and final course grades.

Instructors agreed to enable the AI Tutor for their courses, allowing students to access it when they needed support with individual homework questions. This created a realistic treatment situation in which each student independently chose whether to seek help from the Tutor, and then interact with The tutor to the extent necessary to get the help they desired.

AI TUTOR OVERVIEW

The Tutor is an AI-powered personalized homework tool that supports students whenever they are doing homework assignments within the Achieve platform. It engages students in their homework assignments by guiding them with thought-provoking questions and dialogue, encouraging them to think critically. Designed with research-based guardrails, it helps students make meaningful connections and develop the problem-solving skills that support persistence in their learning journey. Functioning as an extension of an instructor or teaching assistant, the Tutor provides guidance without judgment, removing the stigma of asking basic questions, so students can access support whenever they need it.

STUDY DESIGN

A mixed methods approach, incorporating qualitative insights to complement the quantitative analyses was utilized for this study. This combination allows for a more comprehensive understanding of the findings. The study design uses students as their own counterfactual and the statistical methods and models are rigorous and robust to factors that can potentially bias the results of this study. In short, we will assert that this research provides ESSA Tier 2 evidence of effectiveness. Details of the methodological choices are explained in detail in the following sections.

Ethics and Data Privacy

Prior to data collection, this study and the associated consent forms and instruments were reviewed and approved (found exempt) by the Human Resources Research Organization (HumRRO). HumRRO is an accredited, third-party Institutional Review Board organization with no affiliation with Macmillan Learning. Macmillan Learning seeks third-party review to eliminate any bias in the decision of the exemption. The data in this study, which are provided by the instructor and consenting students, are initially identifiable. However, a random identifier is generated and assigned and then identifiable data are destroyed. Data is stored in secure storage locations, and access is permitted only to the primary investigator in the study.

Sample

The final analytic sample for the student level analyses included a diverse sample of 1,231 consenting college students enrolled in 69 introductory courses (i.e., Calculus, Chemistry, Biochemistry, Biology, Economics, Statistics) across 29 institutions, taught by 37 instructors. Within this sample, students completed a total of 14,495 AI Tutor assignments. Participation required informed consent, ensuring that students voluntarily shared their perspectives, demographic information, course grades, and exam scores. The sample included a range of institutional and course characteristics including variation in enrollment sizes as well as course formats (i.e., face-to-face, virtual synchronous, virtual asynchronous).

Of the 29 institutions, most were four-year colleges (n=19, 66%) with the remainder being two-year institutions offering associate degrees. Most were public institutions (n=23, 79%), and considered large based on enrollment size (n=21, 72%). Nearly half were moderately or more selective (n=13, 45%), while the rest were categorized as less selective.

The majority of courses were taught face-to-face (n=55, 80%), with the remainder being split between hybrid (n=4) and fully online formats (n=10). Course sizes varied widely. The median course enrolled 27 students. Six courses had between 50 and 99 students, four courses exceeded 100 students, and one online course enrolled 2,200 students. It should be noted that this exceptionally large course did not disproportionately influence the student-level analyses. Although it contributed 308 consenting students (from fall 2024), it was confirmed that omitting these students did not meaningfully alter the results.

The variation in participating institutions and instructors enabled a diverse student sample. A breakdown of the analytic sample can be seen in Table 1.

Table 1. Analytic Student Sample Demographic Breakdown

Category	Percentage
First or second year	80%
White	58%
Asian	16%
Hispanic	15%
Black	6%
Female	60%
First in family to college	18%
Eligible for financial aid	32%
Neurodivergent	10%
High school GPA < 3.5/5.0	18%

Methods

After consenting to be part of the research study, participating instructors were given a research study orientation and a training on the Achieve program including its functionality (45 minutes). Instructors mentioned the availability of the AI Tutor to students indicating it was within institution and course policy to use the Tutor as much as they needed or wanted.

Instructors were not required to assign a specific number of Achieve homework assignments, rather they were instructed to use Achieve as they normally would in their courses. The natural assignment and use of the Tutor by instructors and students reflected realistic implementation patterns. While usage was not standardized, the research team observed and documented student engagement with the Tutor throughout the study.

As part of participating in the study, students were invited to complete six surveys over the course of the semester. The surveys collected sociodemographic information, student perceptions of the AI Tutor's use and impact on their learning, as well as their overall experience using the Tutor.

After the semester, instructors shared consenting student exam scores and grades with the research team. The course performance data was merged with Tutor usage data, student demographic information, and opinion data for analysis.

DATA ANALYSIS

This section describes the process for testing the hypothesis that Tutor usage impacts course performance using a complex study design and statistical model. The analysis controlled for both course and student-level characteristics (e.g., assignment difficulty, prior achievement level, etc.) as well as course implementation factors. Methodological considerations such as inconsistent statistical tests (non-normality, heteroskedasticity, dependence, non-random missing data), practical factors (e.g., missing data, reliable and meaningful measurement), and the extent to which the findings generalize to the wider population of students are described.

For this study, Tutor usage was defined as the percentage of an assignment's questions on which a student engaged with the AI Tutor. To examine the effect of usage in a non-linear way, the usage percent was grouped into the following categories: 0%, 1%-24%, 25%-49%, 50%-74%, and 75%-100% of an assignment's questions. The reason for this categorization is explained in the results section.

Assignments were scored as the percentage of questions answered correctly. An assignment had to meet three criteria to be included in the analysis: (1) it was given a score, (2) at least three students in a course completed it, and (3) the scores showed variation greater than 1% (i.e., not all uniformly scored 100% or 0%). Some Achieve assignments did not fit these criteria. For example, a student might open an assignment and receive a 'grade' of 100% or complete an assignment without a grade (e.g., Video, Goal Setting and Reflection Survey, Learning Curve).

Qualifying assignments were further categorized into those with the AI Tutor feature available and those without the Tutor. It should be noted that all students regularly completed their Achieve assignments, either with or without using the Tutor, so student engagement with the assignments is not an issue.

Assignments included in the analyses had a very negatively skewed score distribution (skew = -3.1), that is scores are not normally distributed but rather bunched up near 100%. This is expected with such scores. The mean of the score distribution was 88.6% correct, but the median was 95%. The standard deviation was 17.5%. The 25th percentile was 85% and the 70th percentile reached 100% (i.e., 31% of assignments were scored 100%). To address potential concerns regarding non-normality, robust statistical safeguards were employed to ensure the validity of the results.

One further important distinction regarding Tutor usage is that students have an option to use pre-entered prompts or interact with the Tutor in their own words naturally. Interacting or messaging with the Tutor is how the Tutor was designed and intended to be used by students. The messaging modality is therefore thought to be the most impactful form of interaction. Here students act as their own control, deciding which assignments to use the Tutor, and to use messaging or pre-entered prompts.

The software package IBM SPSS Statistics (Version 28) was used to edit, build, estimate, and provide the results for the statistical models. A complex linear mixed model was used to isolate the unique impact of Tutor use on student assignment (percent questions answered correctly) and course performance (overall course grade and exam average grade as percentages). In order to partial out the unique impact of the Tutor, several factors were included in the model to control for other variables likely to impact academic performance. The variables were:

- course assignment/exam/grade difficulty (equivalent to modeling a random intercept for course)
  - average AI Tutor assignment grade
  - average non AI Tutor assignment grade
  - average course exam score
  - average course grade
- Individual student overall assignment performance (equivalent to modeling a random intercept for student)
  - average AI Tutor assignment grade
  - average non AI Tutor assignment grade
- student prior academic achievement
  - high school grade point average
  - year in college
- student characteristics, economic and home environment factors
  - identify as neurodivergent
  - financial aid eligibility
  - gender
  - ethnic group
  - first person in immediate family to go to college
- course factors
  - number of AI Tutor assignments
  - number of questions on an assignment (assignment level analysis)
  - number of student completed AI Tutor assignments (assignment level analysis)

Including these variables in the model was an attempt to equate students on background variables, prior academic performance, and current academic setting in order to bolster the argument that the observed impact from Tutor usage is not simply a reflection of "better" students using the Tutor more. That is, to the extent possible, all students are equal on all the factors in the model.

Several student characteristics and course level factors were tested to see if they contributed to the model, but they did not. These factors were ultimately not included in the final analyses.

There are additional benefits to the model and methods used here aside from controlling for known and possible confounding factors. Robust (asymptotically consistent) standard errors (i.e., H3 sandwich estimator) were examined. The resulting estimates are robust to the usual bad actors that may bias statistical significance tests or deflated standard errors:

- influential cases (outliers)
- unequal group variances (heteroscedasticity)
- non-normal percent scores
- liberal statistical tests due to nesting of students in courses

Non-response to student demographic questions was modeled by including a non-response category among the other response specific levels of each factor. This method avoids the introduction of bias into the resulting estimates from deleting or improperly imputing data. Modeling non-response allows for explained variation from each student's case for the non-response on each factor, leading to estimates that are unbiased even if the responses are not missing at random. Modeling non-response has the added benefits (over other methods like FIML and Multiple Imputation) of 1.) providing an estimate (when estimable) for the non-responder group on each factor, 2.) overall better control of mediating and confounding factors, and 3.) smaller confidence intervals for estimates.

For example, if the effect for the non-responder group is not statistically significant or practically insignificant, it is likely ignorable, but those incomplete cases could still be included in the analysis for a larger sample size and improved statistical control. This assumption was tested both statistically and practically without additional effort.

On the other hand, if the effect were statistically and practically significant, the full variation was controlled for on that factor and the estimates (direct effects) and statistical tests for important group differences are unbiased by the non-response and the factor as a whole. Only non-response itself is being addressed in this section of text, if however, the cause of the non-response affects the outcome, it should be included in the model and thus controlled for as with any other factor.

Links to the SPSS code used and the statistical output for the student level analyses on exam scores are provided.

- SPSS Code (Link)
- Statistical Tables, Student Level Analyses, Exam Scores (Link)

RESULTS

The findings presented in this report build on earlier results from more than 8,000 students, looking at AI Tutor's role in improving student confidence, engagement and early academic performance. The following results are from the fall 2024 and spring 2025 research effort.

Students Use of the AI Tutor

Nearly seven in ten students used the AI Tutor at least once during the semester, and more than a third used it regularly throughout the semester. Students messaged with the Tutor in 44.5% of assignments when the Tutor was available. Sixty-eight percent of students messaged with the Tutor at some point during the semester, and 40% regularly used the Tutor in the 'sweet spot' (i.e., 25% to 75% of assignment questions). Additional details regarding student use of the AI Tutor are provided in the following sections.

Assignment Level Analysis

Assignment level analyses used the assignment grade (i.e., percent questions answered correctly) for each assignment for every student. This means that if a course had 10 students, and each student completed nine assignments where the Tutor was used by any of those students, the assignment level analyses would have 90 data points.

For this analysis Tutor usage was defined as the percentage of an assignment's questions on which the student engaged with the Tutor. To look at the effect of usage in a non-linear way, the usage percent was categorized into usage levels. Table 2 provides a breakdown of the number of students for each Tutor usage level.

Table 2. Distribution of Students Across Usage Levels: Messaging vs. Pre-made Prompt Modes

Usage Level	Messages	Pre-made
None	(n = 397, 31.5%)	(n = 371, 29.5%)
1% to 24%	(n = 258, 20.5%)	(n = 382, 30.3%)
25% to 49%	(n = 236, 18.7%)	(n = 349, 27.7%)
50% to 74%	(n = 265, 21%)	(n = 110, 8.7%)
75% to 100%	(n = 103, 8.2%)	(n = 47, 3.7)

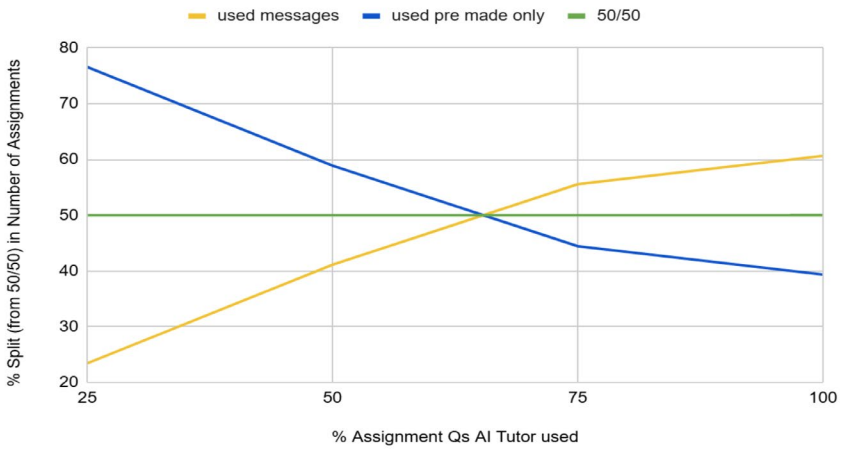
The advantage of categorizing the usage as defined above becomes evident later in this results section when looking at the lines in Figure 5. The gold line is curved, first sloping up with increased Tutor usage, then down again for the highest level of usage. The blue line shows a steady downward or negative trend across levels of usage. The blue line is conducive to a 'linear effect' as it may be reasonable to estimate the trend with a single estimate (for a slope).

In contrast, the gold line is curvilinear, and a 'best fitting' straight line drawn through the points on the graph will underestimate the gain from the Tutor at some points, and over estimate the gain at others. We believe our choice for usage levels allows both situations to be adequately modeled.

Recall that students have an option to use pre-entered prompts or initiate a conversation (message) with the Tutor in their own words. The analysis revealed a contrast in when and how students used these two modalities.

This contrast is presented in Figure 1. Students with lower Tutor usage tend to use the pre-entered prompts more often. At the midpoint of usage the lines cross (i.e., 50% / 50%, green line) and the trend switches to favoring messaging for the highest two usage levels.

Figure 1. Comparative Percent of Assignments Students Messaged With the AI Tutor or Used Pre-made Prompts



Examination of the impact of Tutor usage with messages on assignment grade indicates a 5% to 8% increase over no Tutor use at all (i.e., Table 3). Use of pre-entered prompts has less of an effect from 4% to 6%. Note that these statistical tests have a lot of statistical power because of the thousands of assignments providing the individual data points for this level of analysis.

Table 3. Statistical Contrasts From No Use of AI Tutor With Messages on Assignment Grade

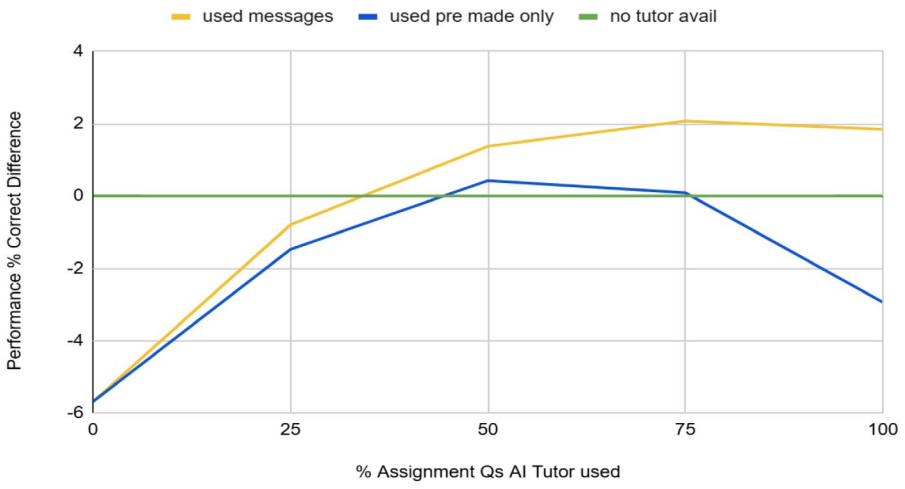
AI Tutor use	Effect	SE	P Value
1% to 24%	4.888	0.783	<0.001
25% to 49%	7.062	0.602	<0.001
50% to 74%	7.761	0.6	<0.001
75% to 100%	7.527	0.654	<0.001
None			

Table 4. Statistical Contrasts From No Use of AI Tutor With Pre-made Prompts Only on Assignment Grades

AI Tutor use	Effect	SE	P Value
1% to 24%	4.21	0.644	<0.001
25% to 49%	6.112	0.591	<0.001
50% to 74%	5.772	0.617	<0.001
75% to 100%	2.748	0.74	<0.001
None			

As evidenced by Figure 2, students who message with the AI Tutor (gold line) can expect a 2% to 5% increase in their assignment grade as compared to students who only use pre-entered prompts (blue line) as usage goes to and beyond 75%. Assignment grade is simply the percent of questions answered correctly.

Figure 2. Student Assignment Performance by AI Tutor Usage



Further, students who message with the AI Tutor (gold line) can expect to produce a modest 1% to 2% increase in their assignment grade over those assignments with no Tutor available (green line).

Student-level Analysis

Student-level analyses were focused on student Tutor use across the semester's assignments. The term 'assignments' refers to Achieve assignments that have the AI Tutor feature present for student use. The results focus on how much Tutor use is necessary for students to see an impact on their exam scores and final grade.

Usage level was defined as the average percent of assignment questions for which a student engaged the Tutor. This was operationalized as a composite score for each student across two interaction modalities: (a) Messaging and (b) Pre-made Prompts. These scores were calculated by averaging the student's engagement across all assignments using a 4-point ordinal scale: 1 (1–24%), 2 (25–49%), 3 (50–74%), and 4 (75–100%).

This approach yields two distinct data points for each student, corresponding to each usage modality. Students are now their own counterfactual, the strongest situation for causal level inference.

A student's level of usage may be different for each modality, but students tended to be similar in the number of questions using messages or pre-entered prompts (i.e., $r = 0.84$ correlation). In addition, the more consistently a student used messages, the less they used pre prompts only.

Figures 3 and 4 indicate a variation in the usage levels across assignments for students. The variation is measured as the standard deviation in the Tutor usage level across a student's assignments. The average standard deviation was 0.44 for pre-entered prompts only, and 0.41 when students used messages. The trend is consistent across modalities and as the number of Tutor assignments completed over the semester increases, with the range topping out at about 1.75.

More than half of the students (57%) using the messaging modality had zero variation or remained in the same usage category for all Tutor assignments. This was true of 34% of students from the pre-entered prompts only modality. These numbers may, however, be exaggerated by including students that did not use the Tutor at all. No use also means no variation in use.

Figure 3. Student Assignment Performance by AI Tutor Usage

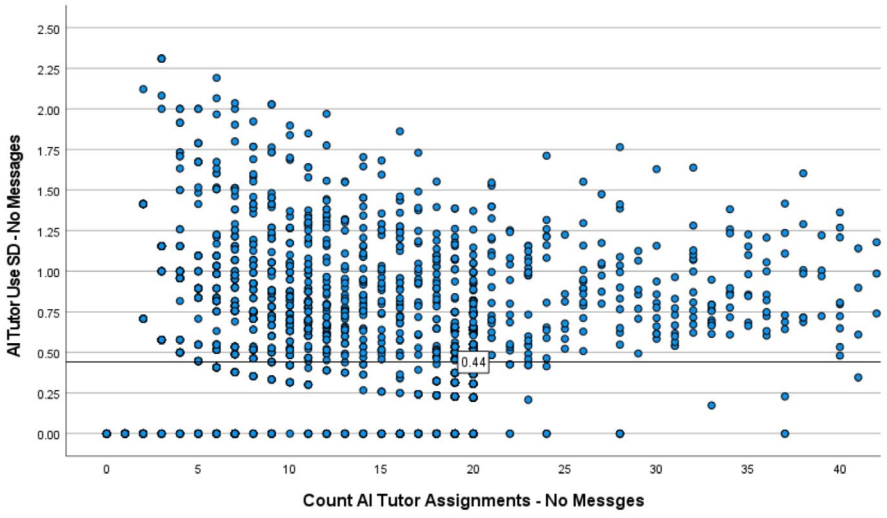
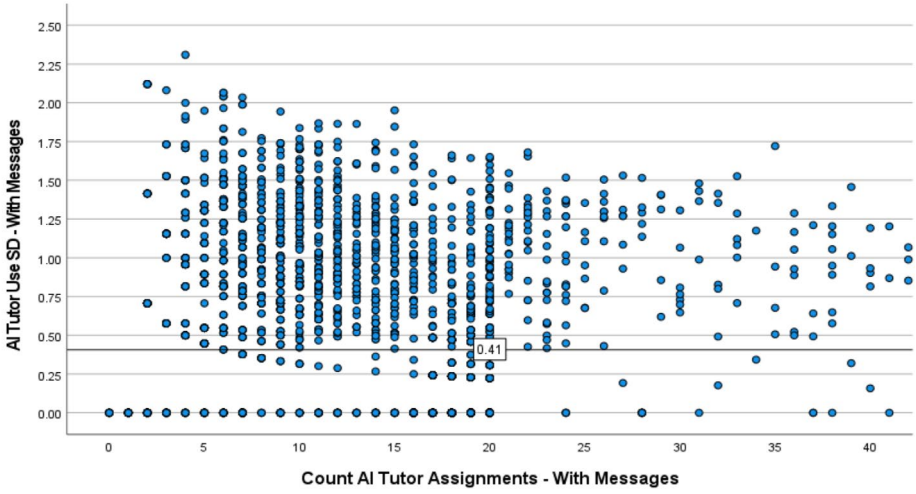


Figure 4. Student Assignment Performance by AI Tutor Usage



When omitting these students, the remaining 2,040 students' assignments with messages had an average standard deviation of 0.88 and a 90th percentile of 1.46. Six percent (6%) remained in the same usage category across all the semester's assignments.

The 2,995 students' assignments with pre-entered prompts only had a smaller average standard deviation of 0.68 and a 90th percentile of 1.16. And again, 6% remained in the same usage category.

As expected the variation increased for both modes when excluding students that did not use the Tutor at all. This finding indicates that most students did not stick to the same usage category, but moved up or down a category from their typical usage. This finding provides evidence students used the Tutor as they deemed necessary.

The next analysis includes the 4,002 students with at least 10 tutor assignments completed (without messages + messages) which represents 87% of enrolled students. More students are represented in this analysis than the analytic sample since there is no need for personally identifiable information nor course performance or student demographic information.

The following results represent the impact of the Tutor on student course performance. Figure 5 displays the impact of Tutor usage on average course exam scores (as a percentage of available exam points). The distribution of exam scores was expectedly moderately skewed (Skew = -0.7) and centered at a C to C+ (Mean = 73.8, SD = 15.8, P_{25} = 64.8, P_{50} = 76.7, P_{75} = 86.1)

Students who message with the AI Tutor can expect to see a 6% to 10% increase in their exam grade, as compared to students who only use pre-entered prompts.

In addition, when using messages students can expect to see a 3.5% to 5.5% increase over no use of the Tutor. Here the routine usage 'sweet spot' is between 25% and 75% of questions (see Table 5). This is a level of usage that seems very doable in practice.

Figure 5. Student Exam Performance by AI Tutor Usage

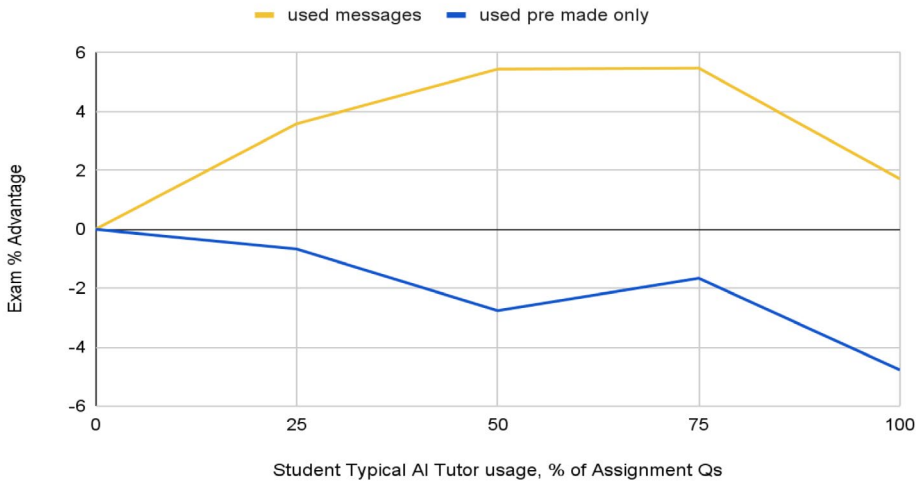


Table 5. Statistical Contrasts from No Use of AI Tutor With Messages (Height of Gold Line) on Average Exam Score

AI Tutor use	Effect	SE	P Value
1% to 24%	3.583	1.16	0.002
25% to 49%	5.439	1.373	<0.001
50% to 74%	5.473	1.315	<0.001
75% to 100%	1.715	1.842	0.352
None			

As the previous figure showed the effect of Tutor usage on student exam scores, Figure 6 displays the impact of Tutor usage on final course grades (as a percentage of available course points). The distribution of final grades was also skewed (Skew = -1.7) and centered at a B to B+ (Mean = 83.2, SD = 12.4, P_{25} = 77.7, P_{50} = 85.5, P_{75} = 91.8)

Students who message with the AI Tutor can also expect to see an increase in their final course grade (2% to 3%), compared to students who only use pre-entered prompts. Additionally there is a 2.5% to 4% increase over no use of the Tutor (see Table 6).

For both exam scores and final grades, the impact dips for the highest message users (i.e., using on 75%-100% of questions). This is, however, only an issue for 8% of students.

Figure 6. Student Final Grade Performance by AI Tutor Usage

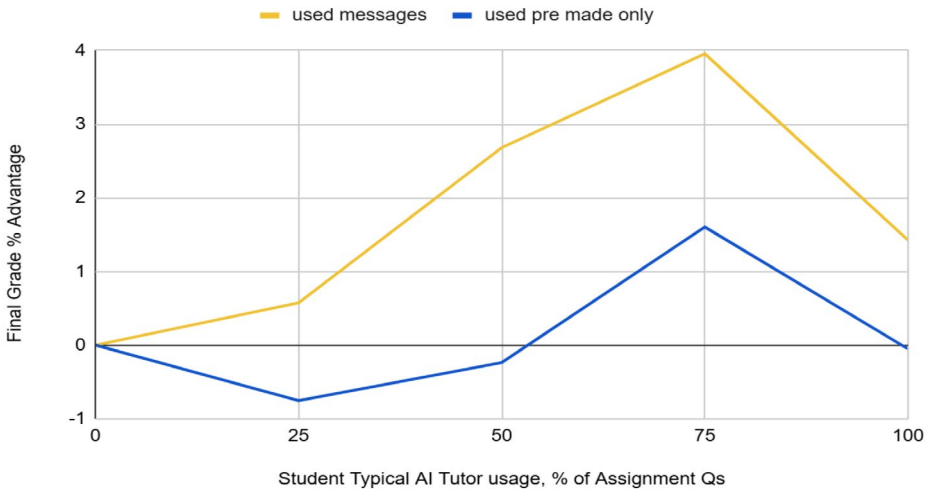


Table 6. Statistical Contrasts from No Use of AI Tutor With Messages (Height of Gold Line) on Final Grade

AI Tutor use	Effect	SE	P Value
1% to 24%	0.575	0.874	0.511
25% to 49%	2.682	1.056	0.011
50% to 74%	3.957	0.961	<0.001
75% to 100%	1.426	1.327	0.283
None			

Student Opinions

Students were surveyed consistently throughout the semesters. A summary of the responses from the spring 2025 semester are shared below. The baseline and post semester surveys had the most student responses with more than 530 responders.

In these surveys, students shared their experience with the AI Tutor [*If you could summarize your experience with the AI Tutor in one sentence, what would it be?*]. Responses fell into two main themes:

1. Transformational and Holistic Learning Support (77% of responses, 398 mentions)
2. Helpful but Inconsistent or Limited (23% of responses, 118 mentions)

Theme 1

Students described the AI Tutor as an essential, helpful, or game-changing learning companion. Comments reflected the Tutor's value in enhancing understanding, building confidence, promoting independence, and generally improving academic performance highlighting the positive impact of the AI Tutor on learning outcomes and study habits.

"The AI Tutor has really helped me better understand my homework and prepare for tests and quizzes."

"Very useful and helpful with understanding material and learning how to solve problems."

"An amazing tool that helped me understand course material better than traditional methods."

"I would say it's very helpful and useful when I'm stuck or need further explanation."

Theme 2

While still generally positive, this theme reflects more nuanced or reserved praise. Students expressed the AI Tutor as helpful at times, but it had limitations. Some described confusion, inconsistent performance, or found it useful only in specific situations. This theme reflects a mixed experience, acknowledging the tool's benefits but also pointing to areas for improvement.

"AI Tutor is a good way to understand topics but sometimes it doesn't give the correct answer."

"A decent aid but it definitely could be better."

"A good help but sometimes confusing."

"Helpful but sometimes it gives different answers than the actual homework solution."

It is important to note that the Tutor does not give the answers regardless of how persistently the student tries. In fact, students complained the Tutor gave them the wrong answer when they thought they tricked the Tutor into giving them an answer. For example, in some cases the student misinterprets the Tutor's feedback when it affirms what the student entered is correct, leading the student to mistakenly assume it is the correct answer to the assignment question. The research team filtered out comments regarding this issue. The Tutor will repeatedly inform the student that it is happy to help, but will not divulge the answer.

Looking at specific questions, Figures 5 through 7, the majority of students feel more confident knowing the AI Tutor is available. For example, students like how the Tutor walks them through the problem, and gives reassurance they are on track. Figure 8 indicates that a majority of students expressed agreement with a variety of positive learning outcomes including:

- Identifying and closing gaps in understanding
- Low stakes practice and test prep
- Critical thinking and reflection
- Independent problem solving
- Homework help anytime when needed

Figure 5. When I have to answer a challenging question, I feel more confident knowing the AI Tutor is there to help.

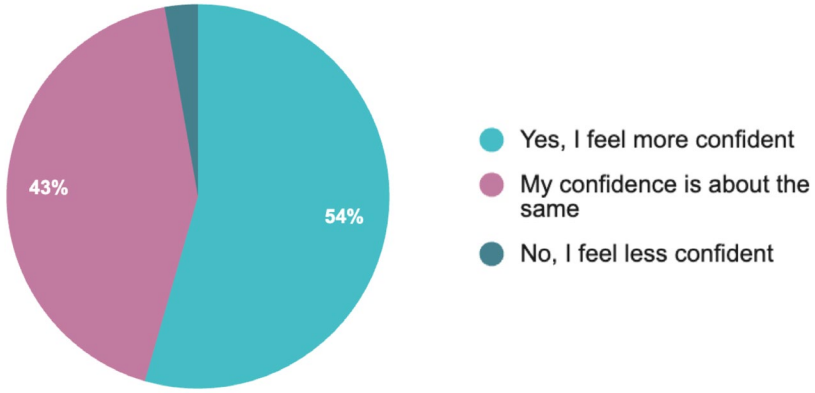


Figure 6. I like that the AI Tutor walks me through the steps of solving a problem, so now I know the steps to solve similar problems.

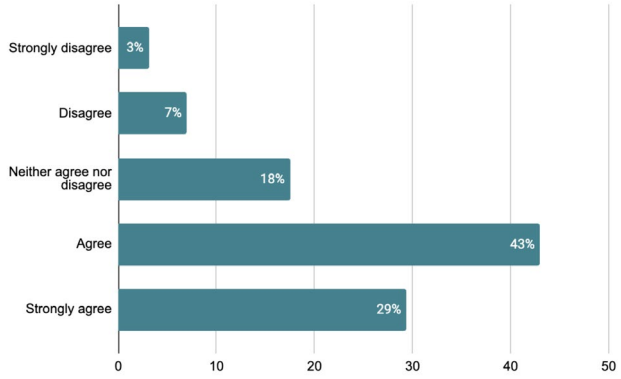


Figure 7. I like that the AI Tutor reassures me that I'm on the right track with a problem.

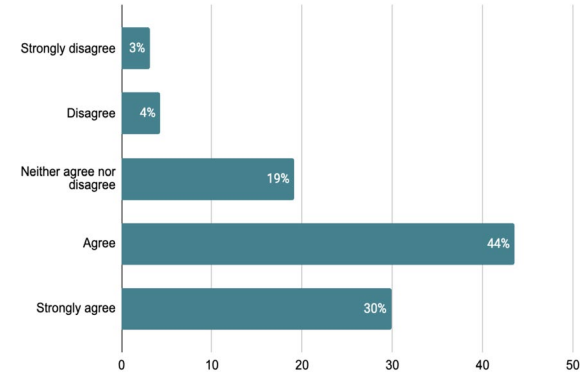
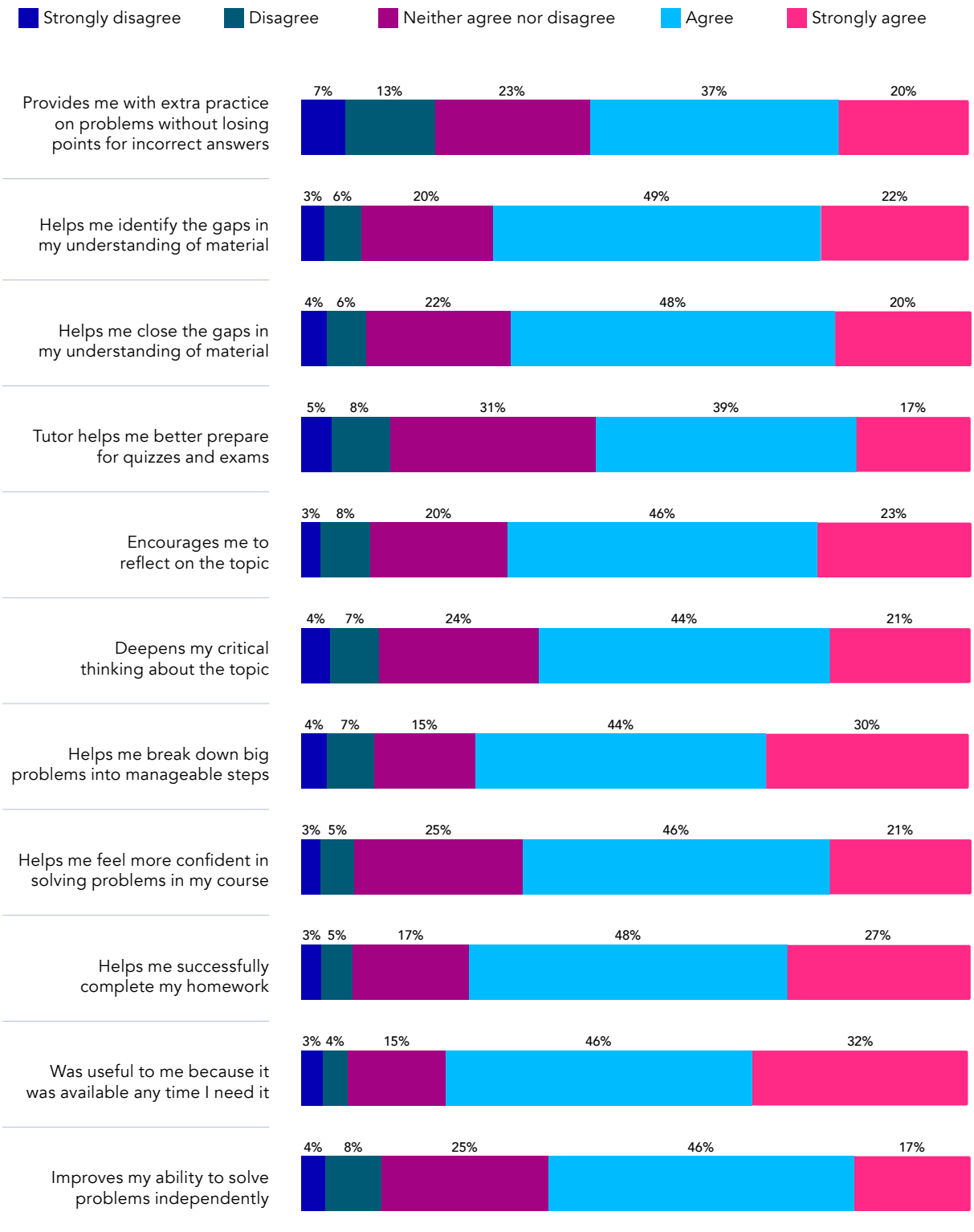


Figure 8. Student Opinions of AI Tutor



IMPLICATIONS FOR STUDENTS AND INSTRUCTORS

Overall, the research findings suggest students have a powerful tool at their disposal in the AI Tutor. The Tutor is always available to provide supportive and personalized assistance on the topic at hand, when the student needs it. This is essential to the learning process.

This finding has important implications for instructors. When students are able to easily seek help for themselves through the Tutor it can reduce the volume of foundational questions instructors receive. As a result instructors can focus their attention on addressing more complex questions.

Students should be encouraged to message with the AI Tutor. The implications from Figure 5 are that many more students could enhance their learning from regular messaging with the Tutor. As a reminder, ideal implementation occurs when students engage with the Tutor on 25% to 75% of the questions.

While the results show students benefit from regular use of the Tutor, the results also suggest that students should not be encouraged to overuse the Tutor (i.e., for every question). Though this was observed for only 8% of the sample, this translates to an average of two or so students in a class. Students should use the Tutor to help them persist, make connections, and think more critically, not necessarily to check every answer. Student feedback indicates that most students agree with this recommendation.

DISCUSSION

The present study investigated the impact of using the AI Tutor in Achieve assignments, across a diverse range of institutions, instructors, and students. While more definitive claims about causality requires additional experimental research, the findings revealed a strong association between regular Tutor usage (i.e., 25% to 75% of questions repeatedly across semester) and students' academic performance (i.e., assignments, exams, grades) while controlling for a range of other relevant factors.

Students are not getting the most out of the AI Tutor if they are not messaging interactively. Analyses indicate that interactive messaging directly impacts their assignment grade, exam grade and final course grade. Consistent interactive messaging with the Tutors provides the optimal learning benefit to students. When conversational interaction with the Tutor is routine, exam scores rise by half a letter grade.

While 40% of students regularly engage with the Tutor through messaging, many more students could benefit from this deeper interaction. Encouraging more use of the Tutor may help students not only receive support when needed, but also explore concepts more deeply.

Lastly, students trusted and like the Tutor, reporting it helped them learn in meaningful ways.

LIMITATIONS AND FUTURE RESEARCH

While the current work represented a large and diverse sample, a convenience sample was used. Students could not be forced to consent, and participate in all data collection procedures. However, the research team did not feel the analytic sample was skewed from the general population on any of the factors measured.

This was not a true experiment with random assignment. A multitude of variables were used to serve as statistical controls, but the lack of random assignment is a limitation. Individual differences of students not captured by the variables used as controls cannot be ruled out as potential confounding variables.

Instructors' implementation of Achieve assignments was also not controlled, and students were free to use the Tutor as much or as little as they deemed necessary. This is a possible limitation of field research conducted in realistic settings. The advantage is that unlike situations where a strict dosage of a treatment can be measured out and received, intervention in educational settings is complex and ever changing, making realistic scenarios valuable research designs. This does however mean that observation, monitoring, and statistical controls become of paramount importance.

Future experimental studies could test the impact of AI tutors by randomly assigning students within the same course or instructor to either receive tutor activities or not. This design would help strengthen arguments of causality by ruling out both individual differences and instructional differences as potential explanations for group differences. This type of research is already planned.

REFERENCES

Doo, M. Y., Bonk, C., & Heo, H. (2020). A meta-analysis of scaffolding effects in online learning in higher education. International Review of Research in Open and Distributed Learning, 21(3), 60-80.

Hurtado, S., & Carter, D. F. (1997). Effects of college transition and perceptions of the campus racial climate on Latino college students' sense of belonging. Sociology of Education, 70, 324-345.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin, 119(2), 254–284.

Lim, L., Bannert, M., van der Graaf, J., Singh, S., Fan, Y., Surendrannair, S., ... & Gašević, D. (2023). Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. Computers in Human Behavior, 139, 107547.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary educational psychology, 25(1), 54-67.

Shute, V. J. (2007). Focus on formative feedback (ETS Research Report No. RR-07-11). Educational Testing Service.