

## Incentivizing supplemental math assignments and using AI-generated hints is associated with improved exam performance

Yifan Lu<sup>1,\*</sup>, K. Supriya<sup>2,†</sup>, Shanna Shaked<sup>2,3</sup>, Elizabeth H. Simmons<sup>4</sup>, and Alexander Kusenko<sup>1,5</sup>

<sup>1</sup>*Department of Physics and Astronomy, University of California Los Angeles, Los Angeles, California, USA*

<sup>2</sup>*Center for Education, Innovation, and Learning in the Sciences, University of California Los Angeles, Los Angeles, California, USA*

<sup>3</sup>*Institute of Environment and Sustainability, University of California Los Angeles, Los Angeles, California, USA*

<sup>4</sup>*Department of Physics and Astronomy, University of California San Diego, La Jolla, California, USA*

<sup>5</sup>*Kavli Institute for the Physics and Mathematics of the Universe (WPI), UTIAS The University of Tokyo, Kashiwa, Chiba, Japan*



(Received 3 January 2025; accepted 27 May 2025; published 12 June 2025)

Inequities in student access to trigonometry and calculus are often associated with racial and socioeconomic privilege, and often influence introductory physics course performance. To mitigate these disparities in student preparedness, we developed a two-pronged intervention consisting of (1) incentivized supplemental math assignments and (2) AI-generated learning support tools in the form of optional hints embedded in the physics homework assignments. Both interventions are grounded in the situated expectancy-value theory of achievement motivation, which posits that students are more likely to complete a task that they expect to do well in and whose outcomes they think are valuable. For the supplemental math assignments, the extra credit was scaled to make it worth more points for students with lower exam scores, thereby creating even greater value for students who might benefit most from the assignments. AI-generated hints were integrated into the homework assignments, thereby reducing or eliminating the cost to the student, in terms of time, energy, and social barriers or fear of judgment. Our findings indicate that both these interventions are associated with increased exam scores; in particular, the scaled extra credit reduced disparities in completion of supplemental math assignments. These interventions, which are relatively simple for any instructor to implement, are therefore very promising for creating more equitable undergraduate quantitative courses.

DOI: 10.1103/v93w-ytyc

### I. INTRODUCTION

Students' mathematical skills play an important role in shaping their learning and performance in undergraduate introductory physics courses [1–4]. However, students enter these courses with a wide variation in their prior exposure to advanced mathematics (i.e., trigonometry and calculus). Moreover, student access to advanced math courses is associated with racial and socioeconomic privilege, so that Black, Hispanic/Latine, and Native American students, and students from low-income backgrounds are

less likely to have had the opportunity to take advanced math courses in high school [5–8].

These disparities in learning advanced mathematics were further exacerbated during the COVID-19 pandemic when students spent an extended time learning remotely [9,10]. Many high school students, especially Hispanic/Latine, Black, and Native American students and those from low-income backgrounds, experienced challenges that made learning more difficult. Some of the most commonly reported challenges included a lack of a quiet space to study or attend class, access to a computer, and a poor internet connection [9,11–14]. As the cohort of students who were in high school during the COVID-19 pandemic enters college, the negative impacts of remote learning on mathematical skills persist and must be addressed to help them succeed in college.

One strategy to address the inequity in prior access to advanced mathematics courses is to provide supplemental instruction to students. This has been found to help students

\*Contact author: yifanlu@ucla.edu

†Contact author: ksupriya@ceils.ucla.edu

with less math preparation successfully complete introductory physics courses [4,15,16]. Most supplemental instruction takes the form of an optional in-person course or series of workshops focused on students' problem solving skills in small groups [15–17]. This can be quite time consuming and resource intensive. Using online learning tools offers an alternative less resource intensive way to approach supplemental instruction but relies more heavily on student motivation and might require students to pay additional fees [4,18]. For example, Forrest *et al.* [4] offered an optional asynchronous online math tutorial to introductory physics students who scored less than 65% on a math diagnostic exam and found that students who completed the online tutorial were 4 times more likely to pass the course. However, this tutorial was not free and only about half of the students completed the online tutorial, even though completing the tutorial was associated with a small amount of course points [4]. Thus, motivating students to participate in optional online supplemental math instruction can be challenging [18,19].

Extra credit points and course credit points are often used to motivate students to complete supplemental math material. Course credit points especially can encourage a larger proportion of students to engage with the optional support offered [18]. However, one potential drawback of offering credit for optional assignments is that it might disproportionately benefit students that are already doing well in the course and might not benefit as much from completing them [18,20,21]. For example, Mikula and Heckler [18] found that when offered extra credit for completing optional math assignments in an introductory physics course, 67% of students with a high grade in the course completed the assignments compared to only 40% of students with a low grade. More concerningly, credit associated with optional assignments might disproportionately benefit students with higher socioeconomic status and fewer responsibilities, who are less likely to need to work full time during college [22]. Scaling the extra credit points so that students with lower scores in the course would benefit more from the optional supplemental assignments offers one way to mitigate this drawback.

An alternative to separate optional supplemental instruction is incorporation of hints within the usual course assignments. Intelligent tutoring systems that students could use for homework have been in place for almost two decades now. These systems have been shown to be effective in improving student learning at the K-12 as well as college level for many disciplines such as mathematics [23–26], accounting [27], and programming [28]. However, the development of these systems was time intensive and required programming expertise. The emergence of large language models (LLMs) such as GPT-4 over the past couple of years has made this much easier [29]. For example, AI-generated hints are offered in all the online homework assignments on kudu.com. The hints are based

on the text of the chapter the students are studying. The effectiveness of such hints generated by AI in supporting student learning is relatively unknown, but it holds much promise.

In this study, we compared the effectiveness of (1) optional assignments containing supplemental math material with scaled extra credit and (2) AI-generated hints on regular course assignments in supporting student learning and reducing inequities in student performance in an introductory physics course at a public university in the Southwestern United States.

### A. Theoretical framework: Expectancy-value theory of achievement motivation

Our interventions in this study are derived from the situated expectancy-value theory of achievement motivation [30,31]. This theory (earlier called expectancy-value theory of achievement motivation) was initially developed in Ref. [32] to understand risk-taking behavior. It was adapted for education by Jacquelynne Eccles and colleagues who were studying achievement behavior shaping sex differences in math achievement among students in fifth through twelfth grades [31]. Over the past few decades, this theory has been widely applied across many educational contexts, including undergraduate education in STEM [33–36]. For example, Perez *et al.* [37] applied expectancy-value theory to understand how students' beliefs in their science competence, task values, and perceived costs may coexist, exploring which combinations may be most relevant for STEM persistence and achievement. According to this theory, students' motivation to engage in an activity is shaped by their expectations of success as well as by the perceived value of the activity [31,38]. In other words, students are more likely to complete a task that they expect to do well in and whose outcomes they think are valuable.

There are four components of achievement values: attainment value (i.e., how important an individual thinks it is to do well on a task for their sense of self), intrinsic value (i.e., enjoyment of the task), utility value (i.e., how an activity is useful for an individual's future goals), and cost (i.e., effort, time, loss of valued alternatives, and perceived cost of failure) (see Eccles *et al.* [31], and Wigfield and Eccles [38] for a more detailed explanation).

In terms of expecting to do well, the AI-generated hints on course assignments could increase student expectations of success on the assignments. In terms of value, scaling extra credit points for the supplemental math assignments could increase the utility value of those assignments, especially for students that did not do well on the midterm exams. And given that the AI-generated hints are integrated into the homework platform in Kudu and do not take much additional time or effort, the cost is minimal.

AI-generated hints also reduce another aspect of the cost of getting help. Students often experience social barriers,

including the fear of negative evaluation, which hinders them from asking instructors, teaching assistants, and/or peers for help [39,40]. However, such social barriers do not exist when using AI tools. AI-generated hints might be particularly helpful for students from marginalized backgrounds that are severely underrepresented in the classroom and/or prone to stereotype threat, which is the term for “when members of a stigmatized group find themselves in a situation where negative stereotypes provide a possible framework for interpreting their behavior, the risk of being judged in light of those stereotypes can elicit a disruptive state that undermines performance and aspirations in that domain [41].”

### B. Research questions

We explored students’ use of optional supports, including supplemental math materials and AI-generated homework hints, and the effects on their exam performance. We were also interested in examining whether these optional supports are used equitably by students with different social identities and thus contribute toward more equitable student outcomes. Specifically, we sought to answer the following research questions (RQs):

- RQ1: To what extent is students’ use of optional supports (specifically supplemental math assignments and AI-generated hints on homework problems) associated with gender, racial, and educational privilege?
- RQ2: Is using optional supports associated with an increase in student exam scores?
- RQ3: Is the use of optional supports associated with student grades similarly for students across a range of social identities?

### C. Positionality of the authors

As researchers and educators, our backgrounds and experiences shape the questions we ask and how we approach this work.

Y. L. is a graduate student in the Department of Physics and Astronomy at UCLA. He is working toward his Ph.D. degree in physics and has been Teaching Assistant for multiple introductory and upper-level undergraduate physics courses. Y. L. identifies as an Asian male.

K. S. is an Associate Director at UCLA’s STEM center for teaching and learning. She is a biologist and STEM education researcher with an emphasis on equity in STEM. K. S. identifies as a South Asian woman.

A. K. is a Professor of Physics and Astronomy with 25 years of experience in innovative teaching. He contributed to development of learning materials and digital tools incorporating AI and enabling *active learning* which are used by tens of thousands of students at UCLA and other universities. A. K. led the development of supplemental mathematics materials and contributed to the development

of AI tools that are used in this study. He was the instructor for the courses included here. A. K. identifies as a White male.

S. S. is a Senior Associate Director of UCLA’s STEM center for teaching and learning. She is a STEM education researcher who has taught introductory physics courses for life sciences majors at UCLA, but has not taught in the series for physical sciences majors described in this paper. In her teaching and education research, she has administered and interpreted the physics Force Concept Inventory. In terms of the identities described in this work, she identifies as a White female.

E. H. S. is a Professor of Physics at UC San Diego with 30 years of experience in interactive teaching. As a long-serving dean and chief academic officer, she has supported inclusive, experiential, and interdisciplinary pedagogy alongside education research across all academic fields. E. H. S. identifies as a White female.

## II. METHODS

The study was carried out at a large public university in the Southwestern United States. The students participating in the study were enrolled in one of two sections (hereafter referred to as sections A and B) of a calculus-based introductory physics course on mechanics, typically the first in a series of general physics courses. Both sections were taught by the same instructor. We obtained demographic data from institutional records. All study protocols were approved by the UCLA IRB. The basic demographic composition of the participants is summarized in Fig. 1. To protect student privacy and identifiable information, we grouped Black/African American, Hispanic/Latine, and Native American students together for analyses because of small cell sizes. Following the recommendation from Asai, we call this grouping “PEERs,” i.e., persons excluded from STEM because of their ethnicity or race, because this term acknowledges that historical exclusion of people based on their race or ethnicity has been systematic and intentional [42].

### A. Study design

There were two forms of optional supports offered in the course: (1) supplemental math materials where students were encouraged to complete practice problems on a set of math skills that are related to mechanics and (2) AI-generated hints for the physics homework questions.

We adopted a quasiexperimental design to assess the association between use of optional supports and student exam performance. All supplemental materials and homework questions assigned to sections A and B were the same. The course was divided into three periods with one exam conducted at the end of each period. While math materials were available to students in both lectures,

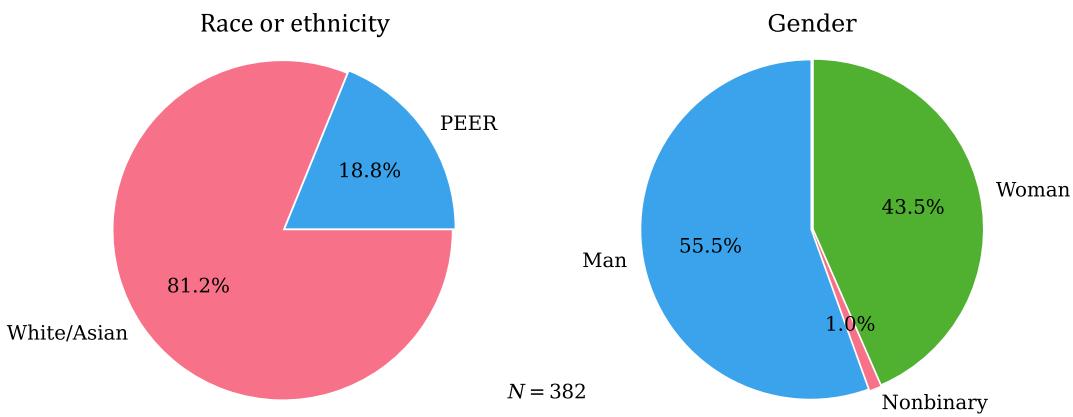


FIG. 1. Demographic composition of the participants in the study. Black/African American, Hispanic/Latine, and Native American students are grouped into the PEER (persons excluded from STEM because of their ethnicity or race) category.

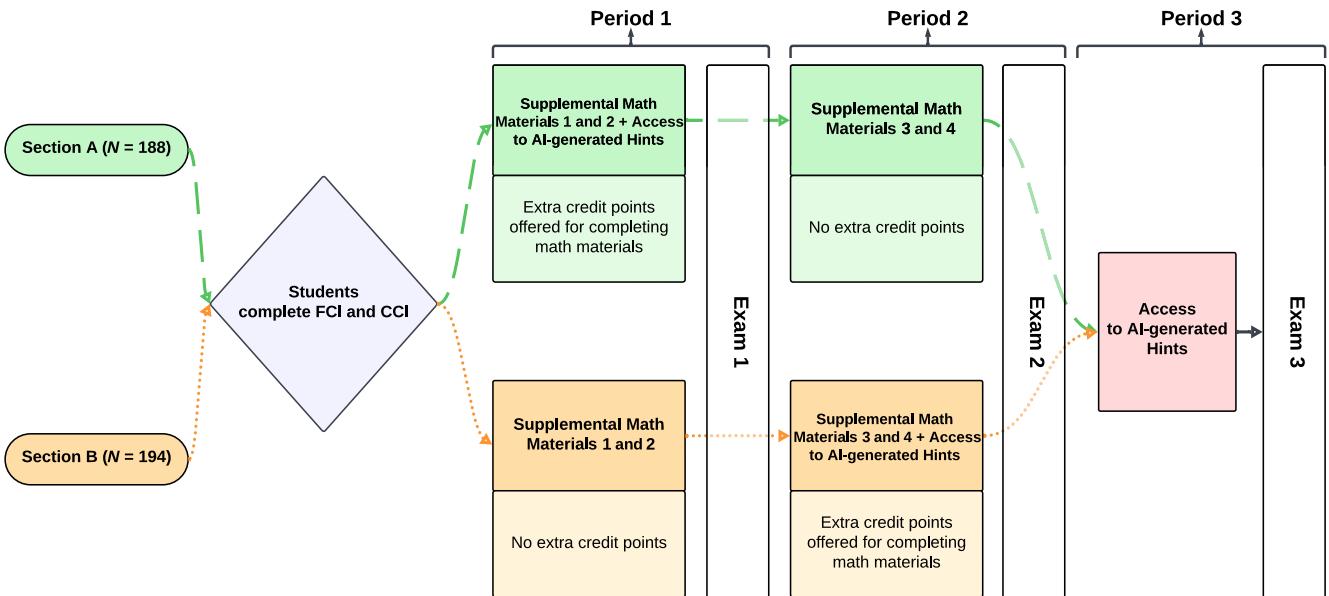


FIG. 2. Research design and the schedule of interventions.

AI-generated hints and extra credit for completing the math materials were only available to one of the two sections during each period (except in period 3). The detailed intervention schedule is listed in Fig. 2.

To measure students' prior preparation, we administered two concept inventories at the beginning of the course to measure students' knowledge of Newtonian mechanics, i.e., the Force Concept Inventory (FCI) [43] and Calculus Concept Inventory (CCI) [44]. Throughout the three periods of this study, we collected data on student homework scores, use of AI-generated hints on the homework problems, and completion of supplemental math materials. We combined this data with demographics to address RQ1 (see Sec. III A). We then performed multiple linear regression with student exam performance as the outcome variable. This allowed us to quantitatively investigate RQ2 and RQ3 (see Secs. III B and III C).

## B. Optional supports

The math skills relevant to students' success in introductory physics courses span a broad range. However, the key components generally involve mastery of vectors and familiarity with calculus-based tools such as derivatives and integration. This was the motivation behind the design of the supplemental math materials adopted in this course, which were divided into four chapters:

- Material 1: Vectors
- Material 2: Derivatives
- Material 3: Integrals
- Material 4: Multiple integrals

Example questions from each topic are shared in Appendix A. At the beginning of the course, the instructor introduced the four supplemental materials together with the possible extra credit points associated with completing them. The materials were then assigned at different stages

of the course. The release schedule was designed to match the ongoing course content while making sure students have abundant time to complete the materials prior to each exam.

Supplemental math materials were accessible to students in both lectures A and B, but incentives for completing them (in the form of extra credit points) were provided at different stages of the course in the two lectures (see Fig. 2). Furthermore, we adopted a formula for applying extra credit so that students who perform better in the exam would gain less by completing the materials. Specifically, the extra credit was determined by both the fractional score  $X$  on the exam and the fractional score  $Y$  on the supplemental material, according to the following formula:

$$\text{Extra credit} = 0.35 Y \cos\left(\frac{\pi}{2} X\right). \quad (1)$$

Note that the due date for the supplemental math materials was prior to the exam, so students did not know exactly how many points the supplemental materials would be worth to them. In other words, students might not have an accurate sense of the utility value of completing the math materials.

The AI-generated hints are qualitative explanations related to physics homework questions and often include key ingredients to solve problems. When a student requests a hint, the AI engine is provided with the text of the entire chapter and the text of the assignment, along with a carefully designed prompt instructing the AI to refrain from giving the full solution to the problem, while pointing the student in the right direction. An example of the AI-generated hints is demonstrated in Appendix B. We have carefully inspected the quality of such hints and made sure that they are both relevant and beneficial in developing students' problem solving skills in this course. When AI-generated hints are available, students have unlimited attempts to use them for each question so that they can explore the most useful version.

### C. Exam question evaluation

There were three exams used for summative assessment in the course, two midterms and a cumulative final. To disentangle the effect of AI-generated hints versus supplemental math material, we evaluated all exam questions to assess their alignment with the math included in the supplemental math assignments and physics homework problems. The evaluation process was carried out as follows: (1) we selected a subset of homework questions that were relevant to the exam questions and created a question pool for each exam. (2) For each exam question, we determined if there was an isomorphic question (i.e., a pair of questions that require the same set of concepts and skills to solve, with differences only at the surface level,

TABLE I. Exam evaluation summary according to the proportion of exam questions that are isomorphic to homework questions, the isomorphism index (ISI), and the proportion of questions that require the same math skills as a given math material, the relevance index (RI).

	RI: ISI	RI: material 1	RI: material 2	RI: material 3	RI: material 4
Exam 1	0.0	0.6	0.2	0.2	0.0
Exam 2	0.4	0.2	0.0	0.2	0.0
Exam 3	0.0	0.8	0.2	0.0	0.0

such as the provided numerical values or the problem context) in the question pool. (3) For each exam question, we also matched it to one or more supplemental math materials according to the math skills required to solve it. An example of a pair of isomorphic questions is given in Appendix C. This evaluation was performed by Y. L., who has a bachelor's degree in physics and has taught undergraduate physics courses.

To quantify our evaluation results, we assigned an isomorphism index (ISI) to each exam, defined as the proportion of exam questions that are isomorphic to homework questions. In addition, for each supplemental math material, we calculated the relevance index (RI) of the exam by the proportion of questions that require the same math skills as a given math material. The evaluation result is summarized in Table I.

### D. Regression models

We used multiple linear regression to assess the association between optional supports and student exam performance. Student exam scores were used as the outcome variable. To standardize the data, we converted the exam scores into z scores with zero mean and standard deviation 1, according to the following transformation

$$z \text{ score} = \frac{\text{raw exam score} - \mu}{\sigma}, \quad (2)$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of each exam. The following variables were used as predictors: (1) prior preparation measured by the concept inventories and converted to z scores, (2) homework scores, (3) optional support usage, including use of AI-generated hints and completion of supplemental math material, and (4) demographics, specifically gender and race or ethnicity. For each exam, we used the question pool from the exam evaluation stage (Sec. II C) and quantified AI-generated hints use as the proportion of problems that a student used AI-generated hints on among all the problems in the question pool. We note that each exam has a separate question pool, so that AI-generated hints use is only measured on the relevant (but not necessarily isomorphic) homework questions. The race or ethnicity and gender

TABLE II. List of variables used in the regression models.

Variable	Description
Homework score	Fractional homework score between 0 and 1
AI hints usage	Fractional problems that students used AI hints on
Math supplemental material	Completion (binary) of materials 1–4
Force Concept Inventory	Force Inventory test score (z score)
Calculus Concept Inventory	Calculus Inventory test score (z score)
Gender	Man (reference) and URG
Race or ethnicity	WA (reference) and PEER

variables were converted to binary variables to protect student privacy given some small cell sizes in the disaggregated data. White/Asian (WA) students were used as reference to PEER (persons excluded due to race or ethnicity including Black/African American, Hispanic/Latine, and Native American students), and men were used as reference to students with underrepresented gender identities (URG: women and nonbinary). The details of all variables used in the model are listed in Table II.

The effect of a variable in the regression model is reflected through its coefficient and its  $p$  value. For RQ3, it is crucial to take into account the effect of using optional support on different student groups. Therefore, we added an additional interaction term in the model, i.e., the product of two variables of interest. For example, we were interested in the interaction between completing math supplemental materials ( $x_{\text{supp}}$ ) and student's race or ethnicity group ( $x_{\text{re}}$ ). So the regression model took the form

$$y = \dots + \beta_{\text{re}}x_{\text{re}} + \beta_{\text{supp}}x_{\text{supp}} + \beta_{\text{int}}x_{\text{re}}x_{\text{supp}}, \quad (3)$$

where the dots indicate regression terms from other variables. When one of the variables in the interaction term ( $x_{\text{re}}$  here) is binary, the interaction coefficient has a simple interpretation: if the binary variable indicates the reference group (taking value 0), then the effect of the second variable is solely expressed in terms of its coefficient. However, when the binary variable takes the value 1, the effect of the second variable comes from the combination of its coefficient and the interaction coefficient. Therefore, in our example,  $\beta_{\text{int}}$  indicates the difference between the effect of completing math supplemental materials between the race or ethnicity groups.

When we tested the assumptions for multiple linear regressions, we found that there is heteroscedasticity in our data. Therefore, to reliably estimate the  $p$  values, we adopted the heteroscedasticity-consistent covariance “HC3” [45] when performing the regression. In addition, some variables in Table II are strongly correlated. As an example, math supplemental materials 1 and 2 are positively

correlated as they are incentivized together. We remedied the multicollinearity issue by dropping some of the variables when necessary (see Sec. III B for more details).

To quantify factors that are associated with students' completion of math materials, we performed a logistic regression. Here, the outcome variable was whether or not students completed a given supplemental math assignment, and predictor variables included prior preparation, demographics, and which section students were enrolled in.

All data analyses were completed using Pandas [46,47], Matplotlib [48], and Seaborn [49].

### III. RESULTS

A total of 188 students in section A and 194 students in section B participated in this study. After converting the FCI and CCI results into z scores, we compared the distribution from section A [FCI: mean = -0.04, standard deviation (SD) = 1.02, CCI: mean = -0.08, SD = 1.04] and section B (FCI: mean = 0.04, SD = 0.99, CCI: mean = 0.07, SD = 0.96). Furthermore, Welch's  $t$  test on the FCI and CCI scores between the two sections showed no significant difference in prior preparation ( $p = 0.41$  and 0.15, respectively). However, we found significant differences in prior physics preparation across gender ( $t = 5.6$ ,  $p < 0.001$ ) and race or ethnicity ( $t = 4.5$ ,  $p < 0.001$ ), and similarly in math background across gender ( $t = 4.1$ ,  $p < 0.001$ ) and race or ethnicity ( $t = 3.4$ ,  $p < 0.001$ ).

Male students scored an average of 0.24 (SD = 1.02) on the FCI compared to URG students who had an average of -0.30 (SD = 0.90). WA students scored an average of 0.10 (SD = 0.99), while PEER students had a mean of -0.44 (SD = 0.91).

On the CCI, male students scored an average of 0.19 (SD = 0.93) while URG students had mean of -0.23 (SD = 1.04). WA students scored an average of 0.09 (SD = 0.98) compared to PEER students who had mean of -0.37 (SD = 1.03).

#### A. Finding 1: Incentives are associated with higher completion rate and reduced disparities in completion of math supplemental material for PEERs

On average, the completion rates for supplemental math assignments when unincentivized were around 20% and offering extra credit increased the completion rate by about 20%. Furthermore, when incentives were not offered, we observed a significantly low completion rate for PEERs in section B, compared to WA students (upper panel in Fig. 3). But this disparity was alleviated when extra credit was offered. Our full logistic regression results are summarized in Appendix D.

We observed no difference in the use of AI-generated hints across gender and race or ethnicity groups, Fig. 4. Most of the students explored the AI-generated hints feature in Kudu: when available, 88% of the students used

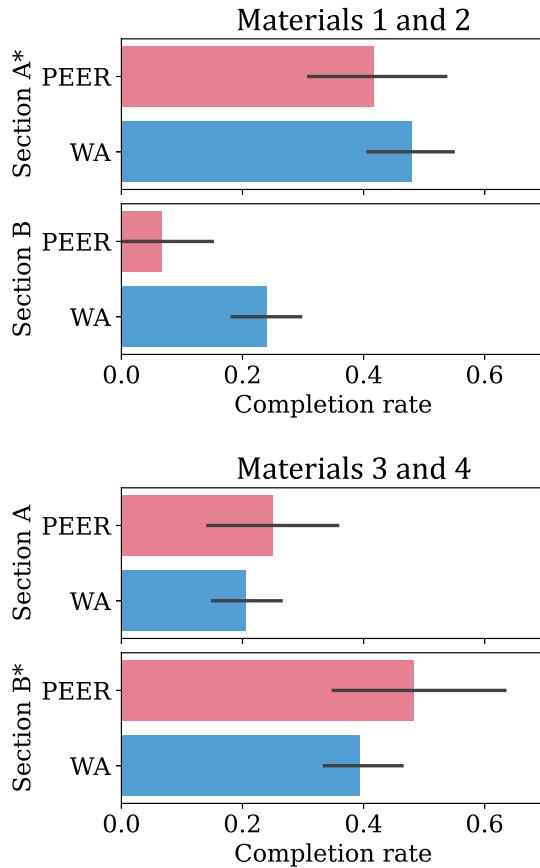


FIG. 3. Comparison of supplemental math material completion rate with different incentive availability across race or ethnicity groups. Stars (\*) indicate incentives are applied to the corresponding section and the black lines mark 95% confidence intervals. PEER stands for persons excluded due to race or ethnicity and includes Black/African American, Hispanic/Latine, and Native American students, WA stands for White/Asian American.

AI-generated hints on at least 1 homework question (among the problems that we found to match the exam questions), and 57% students used AI-generated hints on at least 5 questions. Depending on the section, the question pool size is around 40. On average, students used AI-generated hints on at least 20% of the questions.

### B. Finding 2a: Completing supplemental math assignments was associated with improved student exam performance, especially when math material was most relevant to exam content

The regression model for exam 3 is shown in Table III and we list the results for exams 1 and 2 in Appendix D. None of the exam problems used the math skills covered in supplemental math material 4, thus we excluded this variable from the model. Additionally, there was a strong correlation between math materials 1 and 2 since they were incentivized together, therefore, we dropped the math

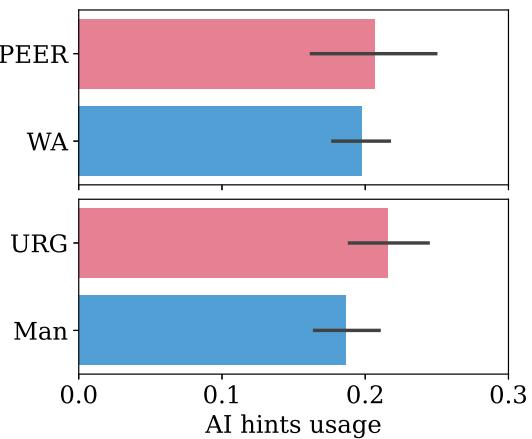


FIG. 4. AI hints usage rate across demographic groups. The black lines indicate 95% confidence intervals. PEER stands for persons excluded due to race or ethnicity and includes Black/African American, Hispanic/Latine, and Native American students, WA stands for White/Asian American. URG stands for underrepresented gender and includes women, nonbinary students, and other gender minorities.

material 2 variable from our models. We did not find a significant correlation between gender and exam performance. However, PEER students had significantly lower exam scores compared to WA students. The correlation between demographics and exam performance was consistent across all exams. In addition, prior preparation in physics and math had a significant association with student exam scores, which is consistent with the findings in Ref. [4].

Notably, the coefficient of supplemental math material 1 completion is positive and statistically significant, showing that additional practice on vectors is indeed positively associated with better exam performance. We did not find a significant association between exam performance and

TABLE III. OLS regression model for exam 3 score. Variables with statistical significance are marked in bold.

	Coefficient	Standard error	p value
<b>Intercept</b>	-2.97	0.34	<0.001
Gender: URG (Reference = man)	-0.04	0.07	0.611
Race or ethnicity: PEER (Reference = WA)	-0.22	0.10	0.030
<b>Homework score</b>	3.16	0.35	<0.001
AI hints usage	0.24	0.21	0.240
<b>Math supplemental material 1</b>	0.24	0.07	0.001
Math supplemental material 3	0.05	0.07	0.444
<b>Force Concept Inventory</b>	0.31	0.06	<0.001
<b>Calculus Concept Inventory</b>	0.12	0.06	0.029

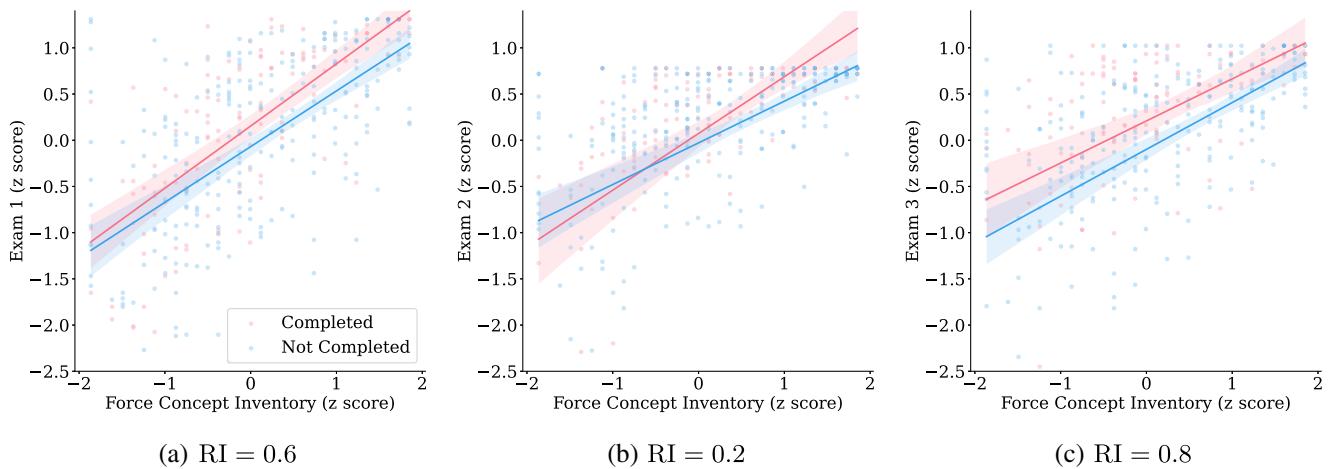


FIG. 5. Correlation of student exam performance with Force Concept Inventory score, separated by supplemental math material 1 completion. The colored bands mark regions with 95% confidence intervals (after taking into account other variables in the model).

other supplemental math assignments. This is because the topics of other supplemental math assignments (derivatives and integrals) are not relevant to the introductory mechanics course we studied. However, those topics are relevant for the subsequent courses in the introductory physics series.

When we compared the relevance of material 1 indicated by our relevance index (RI, i.e., proportion of physics exam questions that used the math skills on a given math assignment) and its regression coefficient across all exams, we found that as RI increases, the regression coefficient becomes not only significant but also larger (see Tables III, VII, and VIII). This effect is visually demonstrated in Fig. 5, where we plotted the best fit lines of exam score versus FCI score for students that completed supplemental math material 1 and those that did not. We observed a clear separation of the fitted lines for exam 1 ( $RI = 0.6$ ) and exam 3 ( $RI = 0.8$ ) with larger separation for exam 3, but this is absent for exam 2, which only has a RI of 0.2.

### C. Finding 2b: Using AI-generated hints is associated with improved student exam performance when exam content is aligned with the homework questions

We found a statistically significant association between exam 2 score and AI hints usage, as demonstrated in Fig. 6 and Table VIII. Interestingly, the linear fit result in Fig. 6 suggests that AI-generated hints might benefit students who are less prepared to a larger extent. To examine this effect, we added an interaction term between the AI hints usage and FCI variables, with the updated regression result shown in Table IX. Indeed, our observation is reflected in the negative coefficient of the interaction term, although not statistically significant. However, we did not find a significant association between student exam performance and AI hints usage for exams 1 and 3, possibly due to the low relevance between these exams and the homework questions, as indicated from the ISI in Table I.

### D. Finding 3: Completing supplemental math assignments might be associated with lower disparities in exam scores between WA students and PEERs

When we added an interaction between competing supplemental math material 1 and race or ethnicity in our regression model, we found that the interaction coefficient was positive with a large effect size (coefficient of 0.38, compared to a coefficient of 0.32 for the Force Concept Inventory) and a  $p$  value of 0.081 (see Table IV). Although the  $p$  value does not reach the conventional threshold at 0.05, the interaction term still suggests that completing supplemental math materials might benefit PEER students more.

The regression coefficients suggest that for two WA students with an average homework score and concept

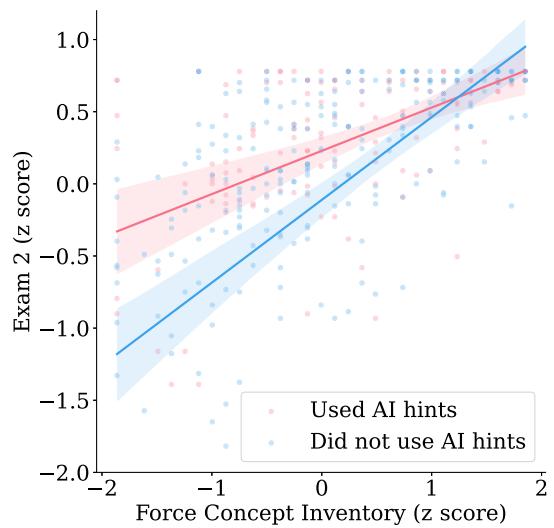


FIG. 6. Exam 2 score versus Force Concept Inventory score, separated by AI hints usage. The colored bands indicate 95% confidence intervals.

TABLE IV. Regression model on exam 3 with interaction term. The last row corresponds to the interaction term. Statistically significant variables are marked in bold and the relevant interaction variables are marked in italics. The positive coefficient for the interaction term indicates that the positive effect of math material 1 is stronger for PEERs.

	Coefficient	Standard error	<i>p</i> value
<b>Intercept</b>	-2.94	0.34	<0.001
Gender (Reference = man)	-0.03	0.07	0.679
<b>Homework score</b>	3.14	0.35	<0.001
AI hints usage	0.24	0.21	0.245
Math supplemental material 3	0.06	0.07	0.414
<b>Force Concept Inventory</b>	0.32	0.06	<0.001
<b>Calculus Concept Inventory</b>	0.12	0.06	0.023
<b>Race or ethnicity</b> (Reference = WA)	-0.32	0.12	0.006
<b>Math supplemental material 1</b>	0.18	0.08	0.020
Race or ethnicity: math material 1	0.38	0.22	0.081

inventory score, the one who completed supplemental math assignment 1 would increase their exam score by about a fifth of a standard deviation. However, for two PEER students who had an average homework and concept inventory score, the one who completed the supplemental math assignment 1 would increase their exam score by more than half of a standard deviation, with contributions from both the math supplemental material 1 coefficient and the interaction term coefficient. Therefore, our result provides some tentative evidence that the additional math practice is not only beneficial to the entire class, but it also improves the exam performance of PEERs' students even more than others.

#### IV. DISCUSSION

We offered two kinds of optional supports to students in an introductory physics course: supplemental math assignments and AI-generated hints. We found that completing the math assignments and/or using AI-generated hints was associated with better exam performance on the exams that were most related to these supports (Figs. 5 and 6).

Previous studies have shown that online supplemental math assignments can be effective in improving student performance in physics courses [4,18]. In this study, we tried a unique scaled incentive structure for free online supplemental math assignments to increase the completion rate, especially among students that need more support with mathematical skills. With this incentive structure, we were able to achieve higher completion rates among students who might need more support in the course compared to students who are well equipped to perform well in the

course, as indicated by students' incoming FCI scores. In other words, when incentivized with the scaled extra credit, about 54% of students that scored below the class mean on the FCI completed the first supplemental math assignment compared to only 37% of students that scored at or above the class mean. By contrast, without the incentive, only 19% of students that scored below the class mean on FCI completed the first supplemental math assignment, compared to 24% of students that scored at or above the mean. When we broke this down by race or ethnicity, we found that our scaled incentive structure was associated with higher completion rate among students from historically underrepresented racial or ethnic backgrounds. The higher supplemental math completion rate among PEERs, combined with the improved performance among students who completed supplemental math assignment 1, suggests that offering these assignments along with a scaled incentive structure can decrease racial inequities in physics courses.

Students who complete math assignments might do better on the exams simply because they are more motivated to do well in the course, and/or because they tend to spend more time working on the course. If motivation and/or time on task were the main reasons for improved exam performance, then we would have seen improved exam performance regardless of the alignment between the exams and math materials. However, our results only show positive association between completing supplemental math assignment 1 and performance on exams 1 and 3, which have the highest proportion of aligned questions; we do not find such an association for exam 2 which had a low proportion of aligned questions. Similarly, we did not find a positive association between exam performance and completing math materials 3 and 4, which were not aligned well with any of the exams in this course. Note that materials 2–4, are useful for downstream courses in the introductory physics series, so students that completed these might perform better in those subsequent courses. We plan to explore this idea in future studies.

Our study is one of the first to explore the impact of offering large language model (LLM)-based AI-generated hints during homework assignments on student performance in exams in an undergraduate physics course. Given the differences in the quality of AI available in open-source compared to paid formats (GPT 3 vs 4 at the time of this study) [50,51], differential access to AI could increase inequities in student performance on homework assignments and exams. In this course, high-quality AI-generated hints for homework problems were made available to all students through the course platform Kudu. We found that a large majority of students used at least one AI-generated hint and there were no significant differences in the use of AI-generated hints by gender and race or ethnicity. Moreover, there was a positive association between use of AI-generated hints and exam 2 performance, which was the exam most aligned with homework assignments.

This suggests that high-quality generative AI tools, if made available to everyone, have the potential to support student learning equitably. In future studies, we plan to explore further students' reasons for using or not using AI hints and their experience with using these hints.

### A. Limitations and future directions

Although our interventions are based on the situated expectancy-value theory of achievement motivation, we did not collect survey data on students' expectancy beliefs and values regarding the supplemental math assignments and AI-generated hints. We also did not collect data on student use of any other AI-powered tools or use of any other optional supports such as tutoring services. We plan to address these shortcomings in future studies. Another limitation of our study is that we were not able to disaggregate the race or ethnicity and gender data because of smaller sample sizes in some of the groups we have included in those categories. We recognize that combining these groups with different historical, social, and cultural context obscures patterns of racial and gender inequality [52]. In future work, we plan to collect larger datasets that could allow us to disaggregate these data for analyses.

A limitation of our intervention is that despite the scaled extra credit incentive structure, the completion rate for supplemental math assignments is still less than 60%. Integrating the supplemental math assignments into formal class time as a component of the Discussion section might further increase student participation. Sharing the results of this study with future students may also be helpful in increasing the perceived utility value of the supplemental math assignments.

In future studies, we also plan to collect some qualitative data using surveys and interviews to give us insights into the factors that shape students' decisions to use the optional supports offered in the course. This qualitative data will help us understand the perceived value of these optional supports and inform structural modifications to increase student uptake of these supports.

### ACKNOWLEDGMENTS

We thank Warren Essey, Stuart Brown, and Joshua Samani for valuable discussions and input. We also thank Harrison Parker for designing the teaching intervention flow chart. This work was supported by UCLA's Teaching and Learning Center and by the UCLA Department of Physics and Astronomy.

### DATA AVAILABILITY

The data that support the findings of this article are not publicly available because of legal restrictions preventing unrestricted public distribution. The data are available from the authors upon reasonable request.

## APPENDIX A: SAMPLE SUPPLEMENTAL MATH QUESTIONS

The supplemental math materials contain both qualitative and quantitative questions covering high school to beginning college level math on vectors, derivatives, and integrals. An example from each material is given below:

#### Material 1 (Vectors):

A two-dimensional vector has an  $x$  component of 8.71 m and a  $y$  component of 5.43 m. Calculate the angle (in degrees) that this two-dimensional vector makes with the positive  $x$  axis.

#### Material 2 (Derivatives):

Calculate the derivative of  $3\sqrt{x^5 + 1} + 5$  with respect to  $x$ .

#### Material 3 (Integrals):

A car antenna that is 0.726 m long (starting at  $x = 0$ ) has a density function  $\rho(x) = e^{x/2}$  kg/m. Find the mass of the antenna. Hint: the mass of an antenna between 0 and  $L$  is  $m = \int_0^L \rho(x)dx$ .

#### Material 4 (Multiple integrals):

A thin disk of radius  $R$  has a uniform mass density  $\rho$ . You are asked to find the gravitational force experienced by a point mass  $m$  located at a distance  $z$  above the center of the disk. Use the formula for the gravitational force due to a mass element  $dM$  at distance  $r$ :  $dF = Gm/r^2 dM$ . Which of the integrals represents the correct answer?

- $F = \int_0^R \frac{2\pi Gmp}{r^2+z^2} r dr$
- $F = \int_0^R \frac{4\pi Gmp}{r^2+z^2} r^2 dr$
- $F = \int_0^R \frac{4\pi}{3} \frac{Gmp}{r^2+z^2} dr$
- $F = \int_0^R \frac{4\pi Gmp}{\sqrt{r^2+z^2}} dr$
- $F = \int_0^R \frac{\pi Gmp}{\sqrt{r^2+z^2}} r^3 dr$

## APPENDIX B: SAMPLE AI-GENERATED HINTS

The AI-generated hints are designed to give qualitative descriptions on the essential steps toward solving a problem. They are meant to *guide* students in the problem solving process instead of completely giving away the answer. A typical example from introductory mechanics is demonstrated below:

**Problem:** “A bird has a mass of 26 g and perches in the middle of a stretched telephone line. Determine the tension when both halves of the line have an angle of  $5^\circ$  from the horizontal direction. Assume that each half of the line is straight.”

**Hint:** “Start by separating the forces acting on the bird into its vertical and horizontal components. Keep in mind that the force due to gravity acts downward and the tension in the wire acts along the wire at angle  $\theta$  to the horizontal. Implement Newton’s second law, which states that the sum of the forces must equal to zero as bird is in equilibrium. Solve the equations to find the tension.”

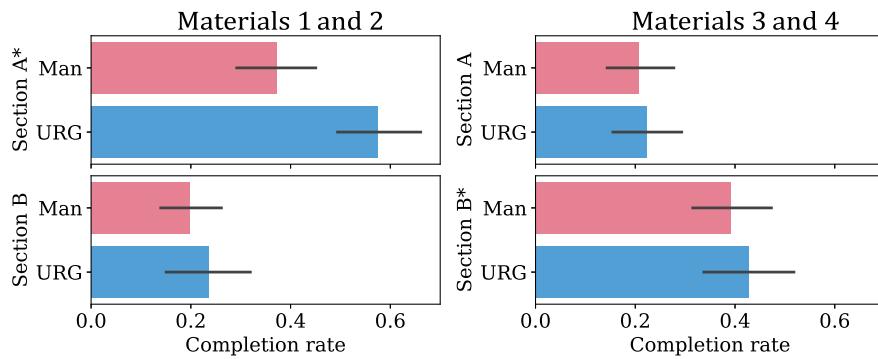


FIG. 7. Comparison of supplemental math material completion rate across gender groups. Stars (\*) indicate incentives are applied to the corresponding section and the black lines mark 95% confidence intervals.

### APPENDIX C: SAMPLE ISOMORPHIC QUESTIONS

In this section, we demonstrate a pair of exam and homework questions that are determined to be isomorphic in our exam question evaluation phase (Sec. II C).

**Homework problem:** “Block 2 slides along a horizontal frictionless table as block 1 falls. The blocks are connected via a frictionless pulley. Find the speed of the blocks after they have each moved 2.0 m. Assume that they start at rest and that the pulley has negligible mass. Use  $m_1 = 2 \text{ kg}$  and  $m_2 = 4 \text{ kg}$ .”

**Exam problem:** “Block 2 with mass  $m_2 = 3 \text{ kg}$  slides along a horizontal table as block 1 with mass  $m_1 = 2 \text{ kg}$  moves downward. The blocks are connected by a rope via a frictionless pulley; the rope and the pulley have negligible mass. Find the speed of the blocks after they have each moved 2.0 m starting from rest.”

### APPENDIX D: OTHER RESULTS IN THE STUDY

Here, we compile the rest of the results in our study not shown in the main text, including:

TABLE V. Logistic regression model for math materials 1 and 2 completion, where the completion variable is 1 if the student completes both materials. Variables with statistical significance are marked in bold.

	Coefficient	Standard error	p value
<b>Intercept</b>	-1.00	0.23	<0.001
<b>Section</b>	-1.16	0.27	<0.001
<b>(Reference = section A)</b>			
<b>Gender</b>	0.77	0.27	0.004
<b>(Reference = man)</b>			
<b>Race or ethnicity</b>	-0.78	0.39	0.045
<b>(Reference = WA)</b>			
Force Concept Inventory	-0.19	0.16	0.243
Calculus Concept Inventory	0.24	0.16	0.131

1. Comparison of math material completion rates across gender groups (Fig. 7).
2. Logistic regression model predicting students’ completion of the math supplemental materials (Tables V and VI).

TABLE VI. Logistic regression model for math materials 3 and 4 completion, where the completion variable is 1 if the student completes both materials. Variables with statistical significance are marked in bold.

	Coefficient	Standard error	p value
<b>Intercept</b>	-2.00	0.27	<0.001
<b>Section</b>	1.06	0.28	<0.001
<b>(Reference = section A)</b>			
Gender	0.01	0.27	0.974
(Reference = man)			
Race or ethnicity	0.14	0.34	0.687
(Reference = WA)			
Force Concept Inventory	0.10	0.16	0.548
Calculus Concept Inventory	-0.18	0.16	0.244

TABLE VII. OLS regression model for exam 1 score. Variables with statistical significance are marked in bold.

	Coefficient	Standard error	p value
<b>Intercept</b>	-1.24	0.37	0.001
Gender	0.01	0.08	0.895
(Reference = man)			
<b>Race or ethnicity</b>	-0.33	0.12	0.004
(Reference = WA)			
<b>Homework score</b>	1.38	0.38	<0.001
AI hints usage	-0.18	0.20	0.337
Math supplemental material 1	0.13	0.08	0.122
<b>Force Concept Inventory</b>	0.42	0.05	<0.001
<b>Calculus Concept Inventory</b>	0.20	0.06	<0.001

TABLE VIII. OLS regression model for exam 2 score. Variables with statistical significance are marked in bold.

	Coefficient	Standard error	<i>p</i> value
<b>Intercept</b>	-3.20	0.31	<0.001
Gender (Reference = man)	0.00	0.08	0.967
<b>Race or ethnicity</b> (Reference = WA)	-0.32	0.105	0.002
<b>Homework score</b>	3.34	0.32	<0.001
<b>AI hints usage</b>	0.40	0.14	0.004
Math supplemental material 1	0.07	0.08	0.360
Math supplemental material 3	0.07	0.08	0.372
<b>Force Concept Inventory</b>	0.29	0.05	<0.001
<b>Calculus Concept Inventory</b>	0.15	0.05	0.003

3. Ordinary linear regression model for exams 1 and 2 performance (Tables VII and VIII).
4. Regression model for exam 2 with additional interaction term (Table IX).

Our logistic regression model predicts the binary variable that corresponds to the completion of supplemental math materials, where the prediction outcome can also be interpreted as the probability that a student will complete the material, given the explanatory variables of the model. In our model, we combined materials 1 and 2 (similarly, for materials 3 and 4) and define “completion” as completing both materials. From Tables V and VI, we can clearly see the effect of offering extra credits on the completion status: the positive (negative) coefficient of the “section” variable in materials 1 and 2 (materials 3 and 4) demonstrates that students are more likely to complete the material when

TABLE IX. Regression model on exam 2 with interaction term. The last row corresponds to the interaction term. Statistically significant variables are marked in bold and the relevant interaction variables are marked in italics. The negative coefficient for the interaction term indicates that the positive effect of using AI hints is stronger for less prepared students.

	Coefficient	Standard error	<i>p</i> value
<b>Intercept</b>	-3.18	0.31	<0.001
Gender (Reference = man)	0.01	0.08	0.875
<b>Race or ethnicity</b> (Reference = WA)	-0.32	0.11	0.002
<b>Homework score</b>	3.32	0.32	<0.001
Math supplemental material 1	0.07	0.08	0.353
Math supplemental material 3	0.07	0.08	0.399
<b>Calculus Concept Inventory</b>	0.15	0.05	0.004
<b>AI hints usage</b>	0.33	0.13	0.008
<b>Force Concept Inventory</b>	0.33	0.06	<0.001
<b>AI hints usage: Force Concept</b>	-0.31	0.18	0.094

extra credits are offered. Interestingly, students’ gender and race or ethnicity play a role in the likelihood of completing math materials 1 and 2; however, they are not significant any more for math materials 3 and 4. This is also reflected in Fig. 3, and such observation could possibly be explained by a temporal effect: the initial incentives help students put a higher utility value on math materials. If students themselves are aware of the benefits of completing the math materials, then they are more likely to complete the math materials even if there is no extra credit associated with it, leading to the relatively high completion rate of materials 3 and 4 for PEERs in section A.

- 
- [1] H. Hudson and W. McIntire, Correlation between mathematical skills and success in physics, *Am. J. Phys.* **45**, 470 (1977).
- [2] P. M. Sadler and R. H. Tai, Success in introductory college physics: The role of high school preparation, *Sci. Educ.* **85**, 111 (2001).
- [3] D. E. Meltzer, The relationship between mathematics preparation and conceptual learning gains in physics: A possible “hidden variable” in diagnostic pretest scores, *Am. J. Phys.* **70**, 1259 (2002).
- [4] R. L. Forrest, D. W. Stokes, A. B. Burridge, and C. D. Voight, Math remediation intervention for student success in the algebra-based introductory physics course, *Phys. Rev. Phys. Educ. Res.* **13**, 020137 (2017).
- [5] D. Battey, Access to mathematics: “A possessive investment in whiteness”1, *Curric. Inq.* **43**, 332 (2013).
- [6] D. M. Bressoud, The strange role of calculus in the United States, *ZDM–Math. Educ.* **53**, 521 (2021).
- [7] C. J. Ackins, Equity in opportunity: Minority student enrollment in advanced, advanced placement, and international baccalaureate English and Mathematics courses in Virginia Public High Schools, Ed.D. dissertation, Virginia Tech, 2022, <http://hdl.handle.net/10919/107829>,
- [8] J. Briones, Racialized tracking: The widening of the calculus achievement gap for black and latinx students, *Ronald E. McNair Scholars Res. J.* **15**, 14 (2023).
- [9] A. R. Golden, E. N. Srisarajivakul, A. J. Hasselle, R. A. Pfund, and J. Knox, What was a gap is now a chasm: Remote schooling, the digital divide, and educational inequities resulting from the COVID-19 pandemic, *Curr. Opin. Psychol.* **52**, 101632 (2023).

- [10] M. Schultz, Impact of the COVID-19 pandemic on achievement gaps in high school: A quantitative analysis of SAT math scores, Ph.D. thesis, Indiana Wesleyan University, 2024, <https://www.proquest.com/docview/3064925036?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses>.
- [11] S. K. Haderlein, A. R. Saavedra, M. S. Polikoff, D. Silver, A. Rapaport, and M. Garland, Disparities in educational access in the time of COVID: Evidence from a nationally representative panel of American Families, *AERA Open* **7**, 23328584211041350 (2021).
- [12] G. M. Francom, S. J. Lee, and H. Pinkney, Technologies, challenges and needs of K-12 teachers in the transition to distance learning during the COVID-19 pandemic, *Tech-Trends* **65**, 589 (2021).
- [13] P. Ong, Covid-19 and the digital divide in virtual learning (2020), <https://escholarship.org/uc/item/07g5r002>.
- [14] E. A. Vogels, A. Perrin, L. Rainie, and M. Anderson, 53% of americans say the internet has been essential during the covid-19 outbreak: Americans with lower incomes are particularly likely to have concerns related to the digital divide and the digital” homework gap”, Pew Research Center, 30 (2020).
- [15] V. B. Meling, M.-A. Mundy, L. Kupczynski, and M. E. Green, Supplemental instruction and academic success and retention in science courses at a hispanic-serving institution, *World J. Educ.* **3**, 11 (2013), <https://eric.ed.gov/?id=EJ1158634>.
- [16] L. A. DeSilva, A. Pullen, and J. E. Hasbun, Enhancing student performance in introductory physics in topics related to electricity and magnetism through the use of voluntary workshops, *Eur. J. Phys.* **39**, 035702 (2018).
- [17] E. Burkholder, S. Salehi, and C. E. Wieman, Mixed results from a multiple regression analysis of supplemental instruction courses in introductory physics, *PLoS One* **16**, e0249086 (2021).
- [18] B. D. Mikula and A. F. Heckler, Framework and implementation for improving physics essential skills via computer-based practice: Vector math, *Phys. Rev. Phys. Educ. Res.* **13**, 010122 (2017).
- [19] S. DeVore, E. Marshman, and C. Singh, Challenge of engaging all students via self-paced interactive electronic learning tutorials for introductory physics, *Phys. Rev. Phys. Educ. Res.* **13**, 010127 (2017).
- [20] M. S. Hardy, Extra credit: Gifts for the gifted? *Teach. Psychol.* **29**, 233 (2002).
- [21] M. A. Harrison, D. G. Meister, and A. J. LeFevre, Which students complete extra-credit work, *Coll. Stud. J.* **45**, 550 (2011).
- [22] K. Supriya, C. Bang, J. Ebie, C. Pagliarulo, D. Tucker, K. Villegas, C. Wright, and S. Brownell, Optional exam retakes reduce anxiety but may exacerbate score disparities between students with different social identities, *CBE-Life Sci. Educ.* **23**, ar30 (2024).
- [23] R. Singh, M. Saleem, P. Pradhan, C. Heffernan, N. T. Heffernan, L. Razzaq, M. D. Dailey, C. O'Connor, and C. Mulcahy, Feedback during web-based homework: The role of hints, in *Artificial Intelligence in Education*, edited by G. Biswas, S. Bull, J. Kay, and A. Mitrovic (Springer, Berlin, Heidelberg, 2011), pp. 328–336, [10.1007/978-3-642-21869-9\\_43](https://doi.org/10.1007/978-3-642-21869-9_43).
- [24] M. Feng, N. Heffernan, K. Collins, C. Heffernan, and R. F. Murphy, Implementing and evaluating ASSISTments online math homework support at large scale over two years: Findings and lessons learned, in *Artificial Intelligence in Education*, edited by N. Wang, G. Rebollo-Mendez, N. Matsuda, O. C. Santos, and V. Dimitrova (Springer Nature Switzerland, Cham, 2023), pp. 28–40, [10.1007/978-3-031-36272-9\\_3](https://doi.org/10.1007/978-3-031-36272-9_3).
- [25] K. Sale and K. Muldner, Learning with an algebra computer tutor: What type of hint is best?, in *Proceedings of the Annual Meeting of the Cognitive Science Society* (2019), Vol. 41, <https://escholarship.org/uc/item/8gv8k0t5>.
- [26] C. Cody, M. Maniktala, N. Lytle, M. Chi, and T. Barnes, The impact of looking further ahead: A comparison of two data-driven unsolicited hint types on performance in an intelligent data-driven logic tutor, *Int. J. Artif. Intell. Educ.* **32**, 263 (2022).
- [27] B. G. Johnson, F. Phillips, and L. G. Chase, An intelligent tutoring system for the accounting cycle: Enhancing textbook homework with artificial intelligence, *J. Account. Educ.* **27**, 30 (2009).
- [28] P. M. Phothilimthana and S. Sridhara, High-coverage hint generation for massive courses: Do automated hints help CS1 students?, in *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '17)* (Association for Computing Machinery, New York, NY, 2017), pp. 182–187, [10.1145/3059009.3059058](https://doi.org/10.1145/3059009.3059058).
- [29] V. Rus, Large language models and their implications for conversational tutors and gift, in *Proceedings of the 12th Annual GIFT Users Symposium*, edited by A. M. Sinatra (2024), pp. 77–86.
- [30] J. S. Eccles and A. Wigfield, From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation, *Contemp. Educ. Psychol.* **61**, 101859 (2020).
- [31] J. Eccles, T. Adler, R. Futterman, S. Goff, C. Kaczala, J. Meece, and C. Midgley, Expectancies, values, and academic behaviors, in *Achievement and Achievement Motivation*, edited by J. Spence (W.H. Freeman, San Francisco, CA, 1983), pp. 75–146.
- [32] J. W. Atkinson, Motivational determinants of risk-taking behavior, *Psychol. Rev.* **64**, 359 (1957).
- [33] T. Perez, J. G. Cromley, and A. Kaplan, The role of identity development, values, and costs in college stem retention, *J. Educ. Psychol.* **106**, 315 (2014).
- [34] K. A. Robinson, Y.-k. Lee, E. A. Bovee, T. Perez, S. P. Walton, D. Briedis, and L. Linnenbrink-Garcia, Motivation in transition: Development and roles of expectancy, task values, and costs in early college engineering, *J. Educ. Psychol.* **111**, 1081 (2019).
- [35] F. Wu, W. Fan, C. Arbona, and D. de la Rosa-Pohl, Self-efficacy and subjective task values in relation to choice, effort, persistence, and continuation in engineering: An expectancy-value theory perspective, *Eur. J. Eng. Educ.* **45**, 151 (2020).
- [36] J. Appianing and R. N. Van Eck, Development and validation of the value-expectancy stem assessment scale

- for students in higher education, *Int. J. STEM Educ.* **5**, 1 (2018).
- [37] T. Perez, S. V. Wormington, M. M. Barger, R. D. Schwartz-Bloom, Y.-k. Lee, and L. Linnenbrink-Garcia, Science expectancy, value, and cost profiles and their proximal and distal relations to undergraduate science, technology, engineering, and math persistence, *Sci. Educ.* **103**, 264 (2019).
- [38] A. Wigfield and J. Eccles, Expectancy-value theory of achievement motivation, *Contemp. Educ. Psychol.* **25**, 68 (2000).
- [39] A. A. Jack, (no) harm in asking: Class, acquired cultural capital, and academic engagement at an elite university, *Sociol. Educ.* **89**, 1 (2016).
- [40] A. Yee, The unwritten rules of engagement: Social class differences in undergraduates' academic strategies, *J. Higher Educ.* **87**, 831 (2016).
- [41] S. J. Spencer, C. Logel, and P. G. Davies, Stereotype threat, *Annu. Rev. Psychol.* **67**, 415 (2016).
- [42] D. Asai, Excluded, *J. Microbiol. Biol. Educ.* **21**, 10 (2020).
- [43] D. Hestenes, M. Wells, and G. Swackhamer *et al.*, Force concept inventory, *Phys. Teach.* **30**, 141 (1992).
- [44] J. Epstein, Development and validation of the calculus concept inventory, in *Proceedings of the 9th International Conference on Mathematics Education in a Global Community* (Citeseer, Charlotte, NC, 2007), Vol. 9, pp. 165–170.
- [45] J. G. MacKinnon and H. White, Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties, *J. Econom.* **29**, 305 (1985).
- [46] T. pandas development team, pandas-dev/pandas: Pandas (2020), [10.5281/zenodo.13819579](https://doi.org/10.5281/zenodo.13819579).
- [47] Wes McKinney, Data structures for statistical computing in Python, in Proceedings of the 9th Python in Science Conference, edited by Stéfan van der Walt and Jarrod Millman (2010), pp. 56–61, [10.25080/Majora-92bf1922-00a](https://doi.org/10.25080/Majora-92bf1922-00a).
- [48] J. D. Hunter, Matplotlib: A 2d graphics environment, *Comput. Sci. Eng.* **9**, 90 (2007).
- [49] M. L. Waskom, seaborn: Statistical data visualization, *J. Open Source Software* **6**, 3021 (2021).
- [50] W. Yeadon and T. Hardy, The impact of AI in physics education: a comprehensive review from GCSE to university levels, *Phys. Educ.* **59**, 025010 (2024).
- [51] L. Ding, T. Li, S. Jiang, and A. Gapud, Students' perceptions of using ChatGPT in a physics class as a virtual tutor, *Int. J. Educ. Technol. Higher Educ.* **20**, 63 (2023).
- [52] E. M. Bensimon, The misbegotten URM as a data point, Center for Urban Education, Rossier School of Education, University of Southern California, Los Angeles (2016).