EdWorkingPaper No. 22-564

Let's Chat: Leveraging Chatbot Outreach for Improved Course Performance

Katharine Meyer
The Brookings Institution

Lindsay C. Page
Brown University

Catherine Mata
Brown University

Eric N. Smith
University of Texas, Austin

B. Tyler Walsh
University of Central Florida

C. Lindsey Fifield
Georgia State University

Michelle Tyson
Georgia State University

Amy Eremionkhale
Georgia State University

Michael Evans
Georgia State University

Shelby Frost
Georgia State University

Eye Eoun Jung
Georgia State University

This study reports on the causal effects of using a non-generative artificial intelligence (AI) chatbot to provide course-specific, proactive outreach and support to students in large-enrollment undergraduate courses. Across both an American Government and Microeconomics course, students randomly assigned to receive chatbot messaging were four percentage points more likely to earn an A or B in the courses. Students assigned to treatment were more likely to complete homework and use supplemental instruction opportunities, which provide evidence that increased course engagement may be driving grade outcomes. We also find suggestive evidence the chatbot reduced the likelihood of students dropping or withdrawing from each course. Treatment effects were generally consistent across student demographics, with the exception of women in Microeconomics, who earned final grades that were seven points higher than women in the control group. The chatbot was well-received by students: 82 percent of students who completed an end-of-course survey recommended its continued use and expansion to other courses. This study provides promising evidence that integrating virtual outreach and communication to students in their college courses can enhance student engagement and learning. It also illustrates the capacity of AI for providing timely responses to students' questions, reducing instructors' time answering common questions and allowing them to devote more time to the students who need it most.

VERSION: November 2024

Suggested citation: Meyer, Katharine, Lindsay C. Page, Catherine Mata, Eric N. Smith, B. Tyler Walsh, C. Lindsey Fifield, Michelle Tyson, Amy Eremionkhale, Michael Evans, Shelby Frost, and Eye Eoun Jung. (2024). Let's Chat: Leveraging Chatbot Outreach for Improved Course Performance. (EdWorkingPaper: 22 -564). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/es6b-sm82

Let's Chat: Leveraging Chatbot Outreach for Improved Course Performance

Katharine Meyer, Lindsay C. Page, Catherine Mata, Eric N. Smith, B. Tyler Walsh, C. Lindsey Fifield, Michelle Tyson, Amy Eremionkhale, Michael Evans, Shelby Frost, Eye Eoun Jung*

November 14, 2024

Abstract

This study reports on the causal effects of using a non-generative artificial intelligence (AI) chatbot to provide course-specific, proactive outreach and support to students in large-enrollment undergraduate courses. Across both an American Government and Microeconomics course, students randomly assigned to receive chatbot messaging were four percentage points more likely to earn an A or B in the courses. Students assigned to treatment were more likely to complete homework and use supplemental instruction opportunities, which provide evidence that increased course engagement may be driving grade outcomes. We also find suggestive evidence the chatbot reduced the likelihood of students dropping or withdrawing from each course. Treatment effects were generally consistent across student demographics, with the exception of women in Microeconomics, who earned final grades that were seven points higher than women in the control group. The chatbot was well-received by students: 82 percent of students who completed an end-of-course survey recommended its continued use and expansion to other courses. This study provides promising evidence that integrating virtual outreach and communication to students in their college courses can enhance student engagement and learning. It also illustrates the capacity of AI for providing timely responses to students' questions, reducing instructors' time answering common questions and allowing them to devote more time to the students who need it most.

JEL Classification: I21, I23, O33
Keywords: college success, college persistence, student support, higher education, academic chatbot.

---

* Meyer: The Brookings Institution. 1775 Massachusetts Ave., NW. Washington, DC 20036 (email: kmeyer@brookings.edu). Page: Brown University (email: lindsay_page@brown.edu). Mata: Brown University (email: catherine_mata@brown.edu). Smith: University of Texas, Austin (email: ensmith@utexas.edu). Walsh: University of Central Florida (email: brendan.walsh@ucf.edu). Fifield: Georgia State University (email: cfifield@gsu.edu). Tyson: Georgia State University (email: mberg2@gsu.edu). Eremionkhale: Georgia State University (email: aeremionkhale@gsu.edu). Evans: Georgia State University (email: mevans33@gsu.edu). Frost: Georgia State University (email: sfrost@gsu.edu). Jung: Georgia State University (email: ejung11@gsu.edu). We gratefully acknowledge Georgia State University and especially Dr. Timothy Renick for their support and engagement with this study. We thank seminar and conference participants at the Annenberg Institute at Brown University, the Brookings Institution, the Association for Education Finance and Policy, the Association for Public Policy Analysis and Management, and the Society for Research on Educational Effectiveness for helpful feedback. We are also grateful to the Mainstay team for their insights and feedback. All errors are our own.

I. INTRODUCTION

Despite documented benefits to college completion, more than a third of students who initially enroll in college do not ultimately earn a credential (Snyder and Dillow, 2015), and inequalities in college completion persist along socioeconomic and racial lines (Cohen et al., 2024; Holzer and Baum, 2017; Kena et al., 2014; Ma, Pender & Welch, 2019). Gaps in college persistence and completion are present even among students with similar academic achievement and preparation in high school (Belley & Lochner, 2007; Kena et al., 2014; Long & Mabel, 2012; Ma, Pender & Welch, 2019). Colleges have invested in a variety of resource-intensive interventions to increase college persistence, including providing students with additional financial aid (Castleman & Long, 2016; Page et al., 2014), enhanced advising (Bettinger & Baker, 2014), or the combination of wraparound advising and financial assistance (Clotfelter, Hemelt, & Ladd, 2018; Scrivener et al., 2015; Weiss, Ratledge, Sommo & Gupta, 2019; Scuello & Strumos, 2024). While many comprehensive interventions significantly increase persistence and degree attainment, not all programs have positive effects, and some of the most promising interventions can be difficult to scale and sustain resource constrained institutions (Sommo et al., 2023).

Increasingly, policymakers, educators, and researchers have focused on how informational barriers and administrative hassle factors have stymied students' pathways to graduation. Completing college requires students to navigate both institutional administrative tasks (e.g., applying for financial aid) and academic tasks within courses (e.g., completing homework). In postsecondary education, several promising interventions have shown that text-based outreach and communication can be a low-cost, easy-to-implement, and effective strategy for supporting educational attainment by guiding students through complex administrative processes. In addition to evidence that text-based communication can facilitate the likelihood students enroll in college (Castleman & Page, 2015, 2017; Castleman et al., 2014; Linkow et al., 2021; Ortagus et al., 2020), the same virtual outreach can also help college students to complete required administrative tasks at higher rates and persist in college (Castleman & Page, 2016; Page et al., 2023).

It is an open question whether institutions can leverage these same communication strategies to improve students' core academic experiences in college. More specifically, can targeted outreach affect academic *inputs*, such as study time and assignment completion, in a way that translates into meaningful *outputs*, such as course performance and retention? In this paper, we report on an effort to implement and experimentally test a text-based chatbot with non-generative artificial intelligence (AI) capability to provide course-specific, proactive outreach and support to students in large-enrollment undergraduate courses. Since 2016, a research-practice partnership between Georgia State University (GSU), external researchers, and Mainstay, a technology company, has collaborated to design, build, and investigate the potential of artificially intelligent virtual communication tools (i.e., “chatbots”) to support students to and through college.[1] GSU uses the chatbot to communicate with students via text message through the persona of Pounce, the university’s blue panther mascot. Experimental studies to date have found that the chatbot communication improved first-year enrollment (Page & Gehlbach, 2017) and the completion of administrative tasks necessary for college persistence, such as handling registration holds and refiling the FAFSA (Page, Meyer, Lee, & Gehlbach, 2023). Given positive effects from these initial experimental studies, GSU has made chatbot communication regarding pre-enrollment and other required administrative tasks standard practice with all students who have opted into receiving text-based outreach. About 86% of incoming students each term opt-in for text-based communication from the university.

In this study, we apply the same chatbot technology within the classroom at GSU with the goal of increasing students’ course engagement, performance and completion. To implement this academic chatbot, we drew on insights from GSU student experiences and prior literature to target courses in which the chatbot had the most potential for impact. We first identified courses with historically high “DFW” rates (DFW refers to a student earning a D, F, or withdrawing from a course). Next, we targeted large enrollment courses and online courses where students had fewer opportunities to connect individually with the instructional team or with peers to form academic support systems. GSU also prioritized courses that were consequential to students’ progression – courses that fulfilled graduation requirements or were gateway courses for popular academic majors. Finally, faculty buy-in was essential to testing the chatbot – while the tool did not require high engagement from faculty members, having faculty willing to experiment with student support initiatives was essential to a smooth implementation.

Given these target parameters, GSU identified two large enrollment, asynchronous online courses in the Political Science and Economics Departments to empirically evaluate the

[1] For more information on Mainstay, see www.mainstay.com.

effectiveness of the academic chatbot via randomized controlled trial.[2] GSU first implemented the academic chatbot in "Introduction to American Government" (hereafter "Government"). Nearly all students at GSU take this course to satisfy a state of Georgia graduation requirement,[3] and the focal section of the course in this study enrolls the largest number of students each term. At the beginning of each intervention term, half of students enrolled in the focal courses were randomly assigned a treatment group, which received 2-3 scheduled, customized text messages each week. This outreach provided general information on weekly assignment due dates, targeted nudges to complete late/missing assignments, and encouragement and invitations to engage with the bot or the instructional team with any questions. Customization of the outreach included both *personalization* (e.g., "Hi FIRSTNAME") and *targeting* (e.g., messages differentiated for students who had a missing assignment versus students who were up to date on coursework). Students could text message the chatbot at any time of day and receive AI-generated responses drawing on a pre-programmed content knowledge base developed in collaboration with university administrators and the instructional team. If the chatbot was not able to find a suitable response in its content knowledge base, those students' messages were flagged for the course TA to review and respond.

Following implementation in Government, we expanded the study to multiple sections of "Principles of Microeconomics" (hereafter "Microeconomics") taught by two different professors. At GSU, Microeconomics is a required course for economics and business majors and is one option for students to satisfy a core curriculum social science foundations requirement.[4] This replication, also conducted as an RCT over two semesters, enabled us to examine whether the academic chatbot yielded similar treatment effects across academic subjects and instructors and to further investigate for which students and in which contexts the academic chatbot affected academic inputs and outputs.

To preview our experimental results, across the two courses, the academic chatbot significantly shifted students' final grades, increasing the likelihood that students earned an A or B by four percentage points. Results were similar across the two courses. In subgroup analyses,

---

[2] We pre-registered the intervention and analysis with the Registry of Efficacy and Effectiveness Studies (REES) for each course under Registry ID 8160 (Government) and Registry ID 13760 (Microeconomics).
[3] GSU students are only exempt from taking "Introduction to American Government" through examination (e.g., Advanced Placement exam scores).
[4] Unlike Government, which all students must pass or test out of, the Microeconomics course is one of 15 possible courses students can select to satisfy the social science foundations requirement.

we find generally similar treatment effects across student demographics, with one notable exception. In Microeconomics, women assigned to treatment earned final grades that were seven points higher than women in the control group; they were 11 percentage points more likely to earn a final grade of an A or B and 10 percentage points less likely to DFW. There were no treatment effects for men. Men and women in the control group performed similarly, thus the treatment effect for women resulted in treated women significantly outperforming treatment and control men as well.

We consider mechanisms through which the chatbot may have increased final grades. We find suggestive evidence that treated students were more likely to complete their homework in Microeconomics and were more likely to attend course tutoring, though effects on these intermediate outcomes vary by student characteristics. We find no evidence of differential assignment completion or performance in Government. Ultimately, on end-of-course surveys, students reported enthusiasm for the chatbot, with 82 percent of respondents recommending its continued use in the course and expansion to other courses at GSU.

Our study makes three important contributions. First, we build on a growing body of evidence on the positive effects that virtual outreach and communication can have on student' completion of essential tasks needed to successfully progress through college, providing support for the hypothesis that such outreach can effectively improve student academic outcomes when integrated into students' courses and delivered by a trusted sender (in this case, from the course professor and/or teaching assistant[5]). While there have been promising studies of virtual outreach, not all applications have yielded significant effects. In one related study, Oreopoulos and Petronijevic (2019) found limited effects of a suite of low-touch psychological, peer coaching, and nudge interventions on college students term grades. While they found improvements to student mental health and increases in reported study time (academic input), these proximal effects did not translate into higher course grades or credits earned (academic output). Pugatch and Wilson (2024) show that email outreach (ostensibly sent from a course instructor) can effectively increase students' academic inputs (e.g., completing additional practice problems), though these efforts also did not translate to increases in course grades /

[5] We refer to the individual monitoring the chatbot and responding to student questions as the "teaching assistant" for ease of interpretation. In practice, that individual was a graduate research assistant who had previously served as the TA for the course. At scale, this role has been filled by a traditional teaching assistant in other courses employing the academic chatbot.

academic outputs. In contrast, Carrell and Kurlaender (2023) found that targeted email messages with assignment reminders, encouragement to attend office hours, and notes about current course performance sent directly from a student's professor led students to perceive their professor more positively. Further, the outreach led underrepresented minority students to earn higher grades in the course and ultimately to graduate from college at higher rates. In another study, adding current course performance information to student communications from faculty – which provided students with regular reminders about their course standing – increased subsequent homework performance (Smith et al., 2018). These varied effects on course performance suggest that message design features – such as the sender or customization to individual student circumstances – are likely to be important factors in the effectiveness of virtual outreach to meaningfully drive intermediary inputs and their subsequent academic outputs.

Second, we highlight the barriers students face navigating administrative tasks within college classrooms and the role of chatbots to improve students' academic outcomes. Extant research shows that course structures affect student performance, with students typically earning lower grades in larger courses and in course taught online. For example, research on in-person college class size finds that students randomly assigned to larger-enrollment classes (compared to smaller-enrollment classes) earn lower grades (De Giorgi, Pellizzari, & Woolston, 2012). Research also shows that students taking online courses earn lower course grades by almost half a grade level (0.44 points on a four-point scale) compared to students taking the same courses in-person (Bettinger, Doss, et al., 2017), and studies leveraging the increase in online courses due to the COVID-19 pandemic similarly find large, negative effects of taking a course online (Bird, Castleman, & Lohner, 2022; Kofoed et al., 2021). Online course taking has increased substantially since the pandemic, with 54% of undergraduates taking some or all of their courses online in 2022-23, compared to 36% of undergraduates in 2019-20 (Goulas, 2024). Given the prevalence of large and online course structures in higher education, despite the documented negative effects of those modalities on student outcomes, our work offers important insights into how personalized, text message outreach can help students navigate structural barriers to college success.

Finally, we advance an understanding of how to incorporate AI response technologies into educational settings provide immediate responses to students' questions, instead of having to wait for a university administrator or course instructor to respond. The state of AI in education is evolving rapidly, moving from initial skepticism and concern to broader acceptance and innovative applications (Bick et al., 2024). Initially, the introduction of technologies like ChatGPT in 2022 raised alarms in the educational sector. K-12 schools and universities responded initially with bans and restrictions, fearing these tools would enable cheating and diminish critical thinking skills among students. However, the narrative is shifting as educators and institutions recognize the potential of AI to revolutionize teaching and learning methods (e.g., Wang et al, 2024). While concerns rightly remain about potential misuses, the time is right for research and development to investigate both the potential and the limitations of AI as a tool for enhancing classroom interactivity, personalizing learning experiences and supports, and aiding students in navigating the administratively complex bureaucracies of educational institutions (Jurenka et al., 2024). Further, beyond the evolution of AI in education, students' relationships with technology have evolved, and testing a mobile-based communication strategy informs how universities can best meet students on their preferred communication platforms. While interventions have long leveraged text messaging for administrative tasks, there has been hesitancy to use a modality perceived as more "informal" for coursework communication. However, research indicates that today's students increasingly rely on mobile devices for schoolwork; Canvas, a common learning management system, reported that in 2024 39% of student assignments were uploaded via a mobile device (Wells, 2024). How students view AI as a complement to existing educational supports is an open area of study. This study contributes valuable insights to the application of non-generative AI to support students' course management skills.

II. INTERVENTION CONTEXT AND DESIGN

Institutional Context

Georgia State University (GSU) is a public, research university in Atlanta, GA that enrolls more than 52,000 undergraduate students. GSU is a minority serving institution with 63 percent of students who identify as Black, Hispanic, or of two or more races, and about 53 percent of GSU students receive Pell grants. GSU has a pooled college completion rate nearly identical to the national four-year institution average, with about 56 percent of students earning a bachelor's degree within eight years of initial enrollment (College Scorecard, n.d.).

Focal Course Contexts

Both the Government and Microeconomics courses were offered as online, asynchronous courses taught by full-time GSU faculty. Both focal courses have long been offered online by the instructors implementing the intervention (i.e., were not shifted online due to the COVID-19 pandemic). In addition to being different subjects, the two courses varied in their learning and assignment structure during the implementation terms. Students in Government read a digital textbook embedded in an adaptive learning platform that quizzed them frequently as they progressed through chapters. Students' course grades in Government reflected performance on these reading quizzes, completion of a visit to a local museum (or alternative assignment), and completion of and performance on 3-4 multiple choice exams (taken asynchronously over a one-week period either at the campus testing center or virtually with a digital proctor). In Microeconomics, students' final grades reflected participation in discussion boards, completion of and performance on practice and evaluative problem sets, and performance on evaluative quizzes. The course did not have a set textbook and instead relied on open access resources. Therefore, we did not have a means of capturing reading engagement metrics similar to those recorded by the adaptive learning platform used in Government.

Business as usual

The business-as-usual level of communication varied between the focal courses. In Government, standard communication from the instructor already included regular, targeted, automated email reminders to students (similar to those tested experimentally in Carrell & Kurlaender, 2023). These emails included reminders about upcoming due dates, encouragement about recent performance, and suggestions for students to meet with the professor when they had failed to turn in an assignment or complete an exam. Some of these messages went to nearly all students (e.g., about 90 percent of students received a message congratulating them on completing the readings for chapter three) while some were targeted to a smaller group with specific course flags (e.g., about 4-5 percent of students received a note from the professor after they failed to submit exam 1 before the deadline). In Microeconomics, instructors also leveraged the automatic email feature, sending reminders about upcoming assignments and targeted outreach to students who had not logged into the course website in five days. Those instructors also encouraged students to engage with them via virtual office hours as part of their standard communications.

All students in the analytic sample (treatment and control) had opted in to receive regular text-based communication from GSU's university-wide retention chatbot. This program sends messages to students about upcoming administrative tasks (e.g., when next semester's registration opens) and targeted, data-informed messages about their enrollment and accounts (e.g., notifying students who have a balance due), in addition to relational messages to provide encouragement throughout the semester and to proactively ask students if they need support or if the bot can connect them with a GSU resource. Prior research found large effects of the retention chatbot on student completion of important college persistence tasks (Page et al., 2023).

Intervention Description

All students enrolled in the focal courses received standard communications from the course instructor and teaching assistant, as described above. Via the chatbot platform, each course instructional team sent treatment students 2-3 scheduled text messages each week (for a total of about 40 messages throughout the semester).[6] These text messages were designed to: (1) provide timely reminders about course requirements; (2) provide customized feedback on each student's individual progress; (3) connect students to course-relevant academic supports; and (4) serve as an additional channel of communication between students and their course instructors. The chatbot messages fell into three broad categories: weekly updates, encouragement messages, and reminder messages. Students received weekly updates every Monday to preview their course tasks and responsibilities for that week. These updates were customized by whether students had completed the previous week's assignments. Encouragement messages were signed by the course TA and were crafted to promote a growth mindset and to invite students to provide feedback on how their semester was going. These encouragement messages were used more frequently in Government. Finally, reminder messages were sent to students as needed (e.g., outreach to students who had not completed an online exam by a given time). When students texted in with a question, the system's artificial intelligence (AI) responded with the closest match response in the system knowledge base. When the system flagged a response with a low probability of response match, the question was then directed to and answered by the teaching assistant to

[6] See Appendix B for a representative set of messages from the fall 2021 implementation in Government.

provide personalized follow-up, as needed. The responses provided by the teaching assistant were then used to update the system knowledgebase.[7,8]

One additional novel feature of the Government course chatbot was a function called #quizme through which students could request a quiz on the course material covered in an upcoming exam. Through #quizme, students could receive and answer a set of multiple-choice questions. For each one, the bot would indicate whether the student answered correctly and/or direct the student to where in the textbook they could read more about the topic and find the correct answer. The bot promoted #quizme in several weekly digests and additional promotional messages. Students could activate #quizme during the two weeks prior to each course exam due date.[9] Since the Microeconomics course did not have exams, GSU did not develop and deploy a #quizme tool for that course.

Intervention Development: Pilot Study

In spring 2021, we launched a pilot study of the course chatbot in the target Government section and distributed messages to all enrolled students who consented to text message communication from the university.[10] The pilot study enabled us to collaboratively develop messages aligned with the syllabus, receive qualitative feedback from students about the chatbot experience, and examine engagement patterns. Students enthusiastically recommended the bot – 90 percent of end-of-course survey respondents recommended that GSU continue the bot for this specific course, and about 84 percent recommended expanding its use to other courses. In addition to receiving student feedback, the spring pilot enabled us to develop a more robust bank of academic chatbot responses and better train the bot to the course structure and context.[11] We also adapted messages based on student feedback and engagement patterns. Most notably, we transitioned to sending students weekly digests customized to their course performance to date

[7] The course TA closely monitored the message interface the two hours following a scheduled campaign, given that most students that replied to the academic chatbot did so shortly after receiving a scheduled message. The TA also checked for flagged messages at least once (and often 2-3 times) each day.
[8] For more information on the technical details of this system, please see: https://patents.google.com/patent/US20180131645A1/en.
[9] The course TA updated the bank of #quizme questions throughout the semester to reflect the chapters covered on the next exam.
[10] Of the 828 students enrolled in the course during spring 2021, 705 had previously opted in for texting from the university and received the pilot messages.
[11] For example, during the pilot the bot would often interpret a question about a course due date as a question about a GSU administrative due date (such as FAFSA or registration). The summer following the pilot, the course TA worked to substantially expand the knowledge base to ensure course-specific questions received a course-specific answer.

(e.g., noting whether they had assignments missing or congratulating those who worked ahead) rather than generic notices of upcoming due dates.[12]

Randomization Design

Each semester of the RCT, we identified all students enrolled in the focal courses who had consented to receive text messages from GSU.[13] We randomized these students to either the academic chatbot treatment condition or to the control group. We separately randomized students enrolled as of the first day of class and a second roster of students who enrolled during the semester add/drop period. As a result of these enrollment patterns, students in the first round of randomization received an additional week of messaging (a welcome message and note about first week assignments) relative to students in the second round of randomization. We account for randomization blocks in our analysis (see analytic model below). We do not remove students from analysis who dropped the course during the add/drop period since dropping the course occurred after treatment began.[14]

Analytic Sample

Across the fall 2021, spring 2022, and fall 2022 academic semesters we randomized a total of 1,568 students enrolled in Government, and during the 2022-23 academic year, we randomized 915 students enrolled in sections of Microeconomics. In Table 1 we report balance on student characteristics between treatment and control students (pooled courses in panel A, Government in panel B, Microeconomics in panel C). We observed no significant differences in characteristics between students in the treatment and control conditions.

In general, student demographics were similar across the two courses. A little over half of the Government sample were women, 45% were Black, 22% were white, 25% were first-generation college students, 57% were eligible for the Pell grant, their average high school grade point average (GPA) was around 3.5, and almost 9% had previously attempted the course and were re-taking it. [15] Overall, about 63% of students were freshmen, though the grade-level

---

[12] Smaller edits included reducing the frequency of emojis and formalizing policies around message length.

[13] We pre-registered the intervention and analysis with the Registry of Efficacy and Effectiveness Studies (REES) for each course under Registry ID 8160 and Registry ID 13760. We estimated a minimum detectable effect size of approximate 0.157 for our main outcomes of interest.

[14] We code students who drop the course as having a zero for all outcomes other than when we examine the effect of treatment on dropping the course. Results are robust to removing students who drop the course during add/drop from the analysis and are, if anything, slightly larger in magnitude than for the overall sample; see Appendix Table 3 for results.

[15] The share of students missing high school GPA values ranged from 7% of students in the fall 2022 semester to 16% in the spring 2022 semester. Most transfer students (~70%) are missing high school GPA. We present high

composition of the course varied considerably across intervention terms, from 43% of the fall 2021 class to 78% of the fall 2022 class.[16] In Microeconomics, about half of the students were women, 52% were Black, 18% were white, 21% were first-generation students, 56% were eligible for the Pell grant, and their average high school GPA was about 3.4. Unlike Government, the Microeconomics courses had a higher share of students re-taking the course – 15%. Historic DFW rates were slightly higher in Microeconomics, resulting in a larger pool of students who might retake the course for a higher grade. While overall about 30% of the Microeconomics analytic sample were freshmen, this varied between 12% of the sample in the fall and 47% in the spring semester.[17] Microeconomics is much more likely to be taken by students who are not in their first semester of college. The instructional team posits this is because the course has a math pre-requisite, which many students may complete their first semester.

III. ANALYTIC STRATEGY

Our primary analytic goal is to estimate the effect of being assigned to receive academic chatbot messaging on course performance, course engagement, and student sense of institutional support. To estimate these effects from the data, we use regression models of the following general form:

$$Y_{irzc} = \beta T_i + \boldsymbol{X}\gamma + \rho_r + \lambda_z + \sigma_c + \varepsilon_{irzc}$$

Where Y_{irzc} represents the outcome for study participant *i* randomized in round *r* and enrolled in term *z* in course *c*, T_i is the indicator for assignment to treatment and is equal to one if the study participant *i* is randomized to the academic chatbot group and zero otherwise. $\boldsymbol{X}$ represents a vector of baseline characteristics for individual *i* (included primarily to explain residual variation in outcomes and to improve precision of estimation as a result), and ε_{irzc} is a random error term. We include fixed effects to account for randomization block (ρ_r), academic term (λ_z), and course (σ_c). In both courses we randomized the roster of students enrolled on the

school GPA summary statistics for students with a valid high school GPA value in summary tables. In our impact analyses including covariates we use dummy imputation and code missing high school GPAs as zero and include an indicator for missing GPA in our models.

[16] See Appendix Table 1 for by-term summary statistics for Government.

[17] See Appendix Table 2 for by-term summary statistics for Microeconomics.

first day of course and then ran a second round of randomization among students who enrolled during the add/drop period; in Microeconomics we additionally blocked randomization by course section. We estimate effects for the course samples overall and for selected subsamples to examine differences by student characteristics and to test for equality of treatment coefficients across subgroups of students.[18]

Data and Measures

Most outcomes come from deidentified course gradebooks, course learning management system records (e.g., student time spent reading the online American Government textbook), and GSU administrative records, provided directly to the research team for analysis. We also aimed to understand the effect of chatbot communication on students' class experiences and perceptions of the instructor. To do so, in Government we added questions in the following domains to an existing end-of-course survey that was directed to students in both experimental conditions: organizational support, self-efficacy, and belonging (adapted from PERTS Ascend and Elevate surveys, see Boucher et al., 2021; Paunesku & Farrington, 2020), instructor expectation (adapted from Smith, 2020), perception of achievable challenge (adapted from Mendes et al., 2007), and novel adaptive expectation scenario items developed for the current study. Appendix C reports the specific attitudinal questions we asked students.[19]

Finally, we included a set of survey items to ask treatment participants specifically about their experience with the course chatbot, including the extent to which they found the communication helpful, whether they read the text messages, whether they knew about and/or used the #quizme function (where applicable), and whether they would recommend future use of the chatbot in this and other GSU courses. As we detail below, two limitations to our survey analysis are low response rates and differential survey participation by student characteristics. Other measures of engagement come from the Mainstay message logs. We code incoming student text messages to identify whether and how frequently students messaged the platform as well as characteristics of their messages (e.g., opt-outs vs. questions).

[18] We do not correct for multiple comparisons precisely because the outcomes in this analysis (e.g., completion of assignments and grade on assignments) are highly correlated but provide distinct perspectives on academic engagement. When comparing heterogeneous treatment effects, we do test for equality of coefficients across different subsamples.

[19] We were not able to field an end-of-course survey in Microeconomics.

IV. RESULTS

Main effects: Final course performance

We first examined the intent-to-treat effect of the chatbot intervention on students' academic course performance. The ITT effect is the most policy-relevant outcome to universities which might use this tool in the future, as it estimates the effect of being initially opted-in to receive chatbot messages. We did not run treatment-on-treated (TOT) analyses, since there is no clear indicator for treatment receipt. The only observable variation in treatment received by students was driven by student opt-out behavior – students who dropped the course or opted out from messaging did not receive the full set of messages, but otherwise students assigned to treatment were slated to receive all messages. We treat dropping the course at any time as an outcome of interest (coding students who drop as having a zero for other outcomes – for example, earning a 0 when calculating final numeric grade). Since chatbot opt-out rates were low, we do not run separate analyses to estimate effects for students to whom the full set of messages was sent.

In Table 2, we report the effects of being assigned to receive chatbot communications on students' final course grade and attainment of positive performance benchmarks (earning an A, earning a B or higher, etc.) and on course completion (whether students withdrew or dropped the course). Earning an A, B, or C ensures students receive college credit for the course and enables degree progress. Other outcomes have different impacts on students' transcripts and degree progress. Students can *drop* a course during the formal add/drop period at the start of the semester (approximately the first two weeks) without penalty. Dropping a course is therefore not an inherently negative outcome. However, since the focal courses in this study fulfill various graduation requirements, limiting drops was an outcome of interest. Students can *withdraw* from the course between the add/drop period and a mid-semester withdrawal deadline. While withdrawals do not affect students' grade point averages, GSU students have a limited number of withdrawals they can take throughout their college careers, and future employers or graduate school admissions officers may look unfavorably on excessive withdrawals. Therefore, reducing withdrawals was a goal of the intervention. Finally, for some students a withdrawal may be preferable to earning a course grade of D or F, which will negatively impact their GPA (and in the case of an F, not count toward the degree credits needed for graduation). Therefore, we separately examine the share of students earning a D or F as well as pooled with withdrawals to estimate the impact of the intervention on overall "DFW" rates.

Across both courses, the largest effects of chatbot outreach were on whether students earned an A or a B in the course. Panel A shows the pooled estimates – students assigned to treatment were four percentage points more likely to earn an A relative to 36% of students in the control group (about an 11% increase) and were four percentage points more likely to earn an A or B relative to 61% of students in the control group (about a 7% increase). Point estimates for these outcomes are nearly identical across courses, though are less precisely estimated when using the smaller samples. Figure 1 plots the density of final numeric grade values (from 0-100+), illustrating the greater density of final grades in the 80-100 range for treated students, primarily as a result of students moving from the C to A/B range.

In the pooled sample reported in panel A, treated students were not significantly less likely to DFW or to drop the course relative to the control group. However, panel A masks variation in course completion by subject. In Government (panel B), students assigned to treatment were 3 percentage points less likely to DFW compared to 18% of students in the control group, with no significant effect in Microeconomics. In Microeconomics (panel C) students assigned to treatment were 3 percentage points less likely to *drop* the course relative to 8% of students in the control group, though as described above drops take place early in the semester. Anecdotally, the Microeconomics course had more assignments due in the early weeks of the course than did Government, and perhaps the intervention helped ease the transition into this heavy Microeconomics workload for treated students.

We then examined *for whom* the course chatbot improved final grade outcomes, testing whether there were differential treatment effects by student characteristics. To illustrate the variance in treatment response, in Figure 2a we plot the treatment coefficients from separate regressions run for different subgroups of students (pooled across courses, including other student covariates as controls) estimating the effect of being assigned to treatment on earning a B or higher.[20] We find significant treatment effects on the likelihood students earn a B or higher[21]

[20] For example, the "first generation" row reports the treatment coefficient and confidence interval for a regression limited to only first-generation students who were assigned to treatment or control in either Government or Microeconomics, including all other covariates (*inter alia*, race, sex, prior academic performance, Pell receipt) in the model.
[21] Results are similar if we look at earning an A as an outcome instead.

across multiple subgroups but little evidence of differential impact (e.g., no evidence that the treatment effect for Black students and for white students are significantly different from each other). We ran formal tests of equality across the full set of outcomes reported in Table 2 and found no evidence of substantially different treatment responses by student characteristics.[22]

However, these pooled effects mask some heterogeneity in how different students responded to treatment in the two different subjects. In Figures 2b and 2c we replicate the structure of Figure 2a but run models separately by subject. In both subjects we again see instances in which a subgroup has a significant treatment response while their counterpart does – for example, in Figure 2b upperclassmen in Government were a statistically significant 7 percentage points more likely to earn a B or higher while the difference between treated and control group freshmen was a smaller and not statistically significant 3 percentage points. However, as with Figure 2a, many of the heterogeneous treatment effects reported in Figures 2b and 2c are not significantly different from other subgroup effects.[23] The notable exception is the treatment effect for women in Microeconomics.

In Table 3 we report the treatment effect in Microeconomics for each of the final grade benchmarks for women and men with the formal test of equality across the two subgroup regressions. We consistently see large effects for women and no effect for men, with these subgroup effects statistically different from each other. In Microeconomics, women assigned to treatment earned final grades that were 7 points higher than women in the control group. They were 11 percentage points more likely to earn a B or higher, 10 percentage points more likely to earn a C or higher, and 6 percentage points less likely to drop the course than women in the control group. We note, however, that it does not appear control group women in this course were performing substantially lower than men (averaging final grades of 68.9 and 70.5 respectively). The treatment effect more than closed that modest gap and resulted in treatment-assigned women substantially outperforming both control and treatment-assigned men.

[22] We run formal tests of equality on each of the main final grade outcomes reported in Table 2 comparing treatment effects for subgroups reported in Figure 3: first-generation vs. continuing generation students, Pell eligible vs. Pell ineligible students, women vs. men, course retakers vs. first-time takers, Black vs. white students, and freshmen vs. upperclassmen. We find across those 42 tests of equality, five are statistically significant: for three outcomes in the comparison between Pell eligible and Pell ineligible and for two outcomes in the comparison between course retakers and first-time takers. Full results available upon request.

[23] In Government we find four instances of non-equal treatment effects across 42 tests of equality – two instances for Pell-eligible vs. Pell-ineligible students and two instances for course retakers vs. first-time course takers. In Microeconomics, except for women vs. men, we find only one instance of non-equal treatment effects across the other 35 tests. Full set of tests available upon request.

Mechanisms: Course deliverables

We next examined mechanisms through which the course chatbot increased final grades, turning first to whether the treatment affected student performance on specific course deliverables. There were no treatment differences in assignment completion or performance in Government, as reported in Table 4.[24] In Microeconomics, however, we do find suggestive treatment effects on assignment completion and performance. As reported in Table 5, in the Microeconomics course students had participation, reading check, and assessment assignments. We considered impacts both on assignment performance (as measured by grades directly reported in the gradebook) and on assignment "effort" or completion. For example, students were graded on their top 10 of 11 assignment quizzes, and we score students as meeting the quiz effort requirement if they complete 10 or more quizzes.

While the Microeconomic chatbot outreach did not affect course participation or the reading checks (practice and assessment "Interactive Tools," cumulatively worth 40% of their final grade), students assigned to treatment were 5 percentage points more likely to complete the minimum number of practice quizzes available (and thus earned practice quiz grades 3.79 points higher, worth 15% of their final grade). Treated students then scored 3.47 points higher on the assessment quizzes (worth 35% of their final grade). Several chatbot messages emphasized the value of completing practice quizzes, and these results suggest one mechanism through which the chatbot may have improved final grades was through encouraging students to complete these discrete, well-defined academic tasks and consequently be better prepared for quizzes.

Mechanisms: Course support take-up

Messaging in both courses highlighted the availability of "supplemental instruction" (SI), a form of course-specific tutoring offered at GSU. SI could be one mechanism for higher course performance if students were more likely to attend SI and, in turn, gain a better understanding of course materials through their sessions. Consistent with this possibility, across courses, students assigned to treatment were about two percentage points more likely to attend SI (Table 6). Overall, SI attendance rates in those courses were low with about 7% of students in Government and 2.4% of students in Microeconomics ever attending SI. These low baseline rates make a two-

[24] Sample size for assignments vary because the professor changed assignments across the intervention terms, dropping the fourth exam after the first intervention term, dropping the field trip after the second intervention term, and switching textbooks after the second intervention term to a system that did not track reading time.

percentage point increase large in relative terms. In Figure 3 we report subgroup treatment effects on SI attendance. We find statistically significant increases in SI use among female, Black, upperclassmen, Pell eligible, and continuing generation subgroups. Overall, we take this as suggestive evidence that treatment students were more likely to attend SI, a plausible mechanism through which they may have done better in the course.

Mechanisms: Student attitudes

We next examined impacts on students' end-of-course attitudes, which we measured based on responses to an end-of-course survey. We hypothesized that the chatbot may work to improve students' final grades through improving their sense of connection with the course and academic community. Our survey analysis is limited in several ways. The end-of-course survey was only fielded in Government, completion of the end-of-course survey was voluntary, and across semesters about two-thirds (68 percent) of students completed the end-of-course survey. While treated and control students were equally likely to complete the end-of-course survey, there were stark demographic differences between those who did and did not complete the survey. For example, the shares of Asian students and Hispanic students were higher in the survey completion sample, and the shares of Black students, course re-takers, and students with below-median high school GPAs were significantly lower in the survey completion sample.[25] Thus, the survey results are not necessarily generalizable to all students in the course.

We created indices of grouped measures (e.g., averaging Likert-scale responses across three items that asked about sense of social belonging) and used that category average as our outcome of interest in our main regression model. Among students who completed the survey, we find no consistent treatment effects (see Table 7). Thus, we have little evidence that changes to student *attitudes* (at least on the constructs we measured) were a key mechanism through which the academic chatbot improved student grades. We note this as a potential area for future research with more comprehensive data coverage.

Transfer of chatbot impact into other academic domains

In our final impact analysis, we explored whether the chatbot affected students' outcomes outside of the focal intervention courses. We did not have a prior hypothesis about the impact of treatment on performance in other courses. On the one hand, if students have finite study time available and chatbot messages directed them to spend more time on the focal course, then we

[25] See Appendix Table 3 for full summary of survey completion demographics.

might observe a negative treatment effect on their grades in other courses. On the other hand, if the treatment helped students develop better time management skills or take up needed supplemental supports at higher rates, for example, they may have leveraged those skills to navigate all of their courses and had higher performance overall during the intervention term. If these potential mechanisms operate in different ways for different students, then, on average, we may see little impact on students' performance overall.

In Table 8 we report on students' overall term GPA during the intervention term, their intervention term GPA excluding the focal (Government or Microeconomics) course, and whether they enrolled in another course in the subject the following term. We do not find strong evidence of a positive or negative effect on academic outcomes beyond the intervention course for the overall sample. We did find that students in Government earned higher overall term GPAs (including the focal course in calculations), and in Figure 4 we show this may have been driven by differences between treated and non-treated first-time course-takers and non-Pell students in Government. In Microeconomics, treated students who were both Pell-eligible and first-generation students earned higher overall semester grades than similar students in the control group.[26] An open question remains regarding the extent to which an academic intervention in one course might affect students' long-term academic engagement and performance and the components necessary to affect such change.

Descriptive Analysis: Student Bot Experience and Engagement

In Figure 5 we summarize feedback from the Government end-of-course student survey in which we asked treated students to complete a feedback form on their experiences with the chatbot. About 90% of treated students who completed the end-of-course survey recalled receiving chatbot communication, with 72% of students reporting that they read most of the messages, and 64% reporting that the weekly messages were helpful. While 65% of students said they knew the #quizme tool was available, only around 38% of students reported using #quizme. About a third (35%) of all respondents, representing 89% of the #quizme users, said that the #quizme tool was helpful. That such a high rate of tool users found it helpful, but only half of students aware of the tool reported using it indicates opportunities for future work to explore how students decided whether to use academic supports and whether additional messaging could

[26] We found no overall, by-course, or by-subgroup treatment effects on course-taking next term, results available upon request.

more effectively increase #quizme take-up. When asked whether GSU should continue using the chatbot, 82% indicated that the chatbot should continue being used in Government and that GSU should expand it to other courses.

Finally, in Table 9 we present measures of student engagement based on de-identified logs of all messages exchanged between students and the chatbot platform (inclusive of pre-scheduled messages, automated bot responses, and supplemental human responses). With these data, we can examine measures of treatment dosage, such as how many messages students receive, as well as opt-out rates. The treatment was implemented as intended – over 99% of students assigned to treatment were sent at least one message.[27] Dosage was similar across courses – students received about 44-46 messages throughout the semester, inclusive of scheduled bot messages and responses to their inquiries. Relatively few students opted out – 4% in Government and 3% in Microeconomics.[28] About half of the students ever messaged back to the bot, with slightly higher reply rates in Government (54%). The average number of replies was higher in Government – an average of 4.6 replies overall (and 8.5 replies among the students who ever replied). This may be due to the use of the #quizme feature in Government which was designed for multiple back-and-forth messages as students attempted the sample quiz questions, while Microeconomics had fewer interactive components (about one message on average, or 2.5 messages on average among students who replied at least once). However, students varied considerably in their active engagement – the highest engagers in Government and Microeconomics sent 76 and 24 messages, respectively, throughout the semester.

V. DISCUSSION

In this study, we evaluated the effect of a course-specific academic chatbot providing students with customized, timely, and regular notifications about course requirements and feedback on their performance in large, online sections of undergraduate courses. Given prior work showing that chatbots can successfully improve students' completion of administrative college tasks, we hypothesized that the course-specific integration of chatbot communication

[27] This is not 100% because some students may have dropped the course between randomization (typically conducted the Friday prior to the first week of classes) and the first launch message being distributed (typically the Monday of the first week of classes).
[28] We code opt-outs based on student replies including use of formal opt-out language (e.g., "#pause") and informal requests (e.g., "stop txting me").

would improve overall course performance as well as completion of course tasks, such as completing the readings or turning exams in on time. Particularly given low national college completion rates and completion rates at our partner institution, we hoped the chatbot would support students' short-term course performance and potentially have subsequent effects on persistence and future college engagement.

We find compelling evidence that the chatbot communication shifted students' final course grades, increasing the likelihood that students would earn an A or B in the course. We also find suggestive evidence that the chatbot encouraged course completion, with treated students less likely to DFW in Government and treated students less likely to drop the Microeconomics course. Our heterogeneous treatment effect analyses highlight for whom the bot may be most effective, often indicating larger treatment effects for students facing more substantial barriers to engagement. For example, there is a large, longstanding literature highlighting the underrepresentation of women in the economics profession in general and the importance of diversifying undergraduate economics departments to attract a more representative student body (Bayer & Rouse, 2016; Bayer, Bruich, Cherry, & Housiaux, 2020; Dynan & Rouse, 1997; Yellen, 2019).

Women are underrepresented in college majors such as economics and science, technology, engineering, and math (STEM) in part due to a lower sense of belonging as well as their lower persistence when they receive lower early college course grades (Allen & Robbins, 2008; Good, Rattan, & Dweck, 2012). Further women frequently underestimate how well they are doing in a course or on tasks that are seen as stereotypically male (Coffman et al., 2024). We found large treatment effects of the chatbot on final grades for women in Microeconomics, though we did not find evidence of increased short-term persistence in taking economics courses (measured in the next semester following interventions). These findings suggest course chatbots – or other means of regular and proactive communication– are a potentially promising strategy to increase representation in economics, as students receive both up-to-date information about their course performance and encouragement to connect with campus supports, though more work is needed to understand longer-term effects. We also note that in our study the Microeconomics course was taught by two female instructors, opening the possibility that an interactive effect between messaging and instructor characteristics is driving the heterogeneous treatment effects we observe.

The AI chatbot technology enabled the instructional team to provide targeted, clear information to students about their course performance to date and the necessary tasks to complete to ensure success in the course. It is worth underscoring that the piloting and implementation of the technology required substantial upfront investment from the course instructional team and the university support office. Piloting the academic chatbot for a semester enabled the team to develop messages aligned with the course syllabus and provided time to train the bot on course-specific questions students might ask (as well as time to set up a course-specific #quizme question bank). In addition to targeting courses where students would likely benefit from the academic chatbot (e.g., high enrollment courses, virtual courses, courses with high "DFW" rates), the chatbot was also easier to launch in well-established courses with solidified course syllabi and schedules. Ultimately, we were encouraged by students' positive response to the chatbots, with high awareness of the tool and endorsement of GSU's use of course chatbots to support learning across subjects. Indeed, based on that feedback and this analysis, GSU has adopted the academic chatbot as a status quo tool in these courses.

Successful implementation also requires ongoing human monitoring of incoming messages to ensure students receive timely and accurate responses to their questions. For example, in one message, a student noted they had been dropped from the university (and course) for tuition non-payment. The chatbot replied immediately with the phone number of and a website link to student financial services. In addition, the human teaching assistant for the course was able to follow-up with a note that the professor would be able to provide the student with access to the course textbook while the student resolved their account hold so they would not fall behind on the reading. Successful implementation of a course chatbot requires sustained commitment and attention from the instructional team to ensure a high-quality student experience that is well aligned with the course itself. Notably, after the initial pilot period, weekly time spent monitoring and responding to messages declined substantially, with the course TA typically devoting less than two hours per week to system monitoring.

The exchange regarding tuition non-payment also highlights the importance of providing students with multiple communication channels to reach their instructional team and the pros and cons of AI-enabled messaging. Some students may feel uncomfortable discussing sensitive topics – such as being dropped for account nonpayment – in person with an instructor, but when prompted about a task, may feel more comfortable sharing such information via text message and ultimately receiving a response that facilitates their navigation of challenges they are facing. In this sense, the academic chatbot may support students' sense of psychological safety by offering another channel through which to develop positive relationships and establish trust (Wanless, 2016). Many factors hinder students from seeking help in introductory and first-year courses, including a lack of confidence and uncertainty about how to approach an authority figure such as a college professor (Stitzel & Raje, 2022). A chatbot provides a low-stakes way of asking questions. The chatbot can be a useful source of information that students can access immediately and free of concerns about judgment (de Gennaro et al., 2020). On the other hand, there has been some concern with individuals, especially youth, over-anthropomorphizing AI tools and developing unhealthy parasocial relationships with technology (Toppo, 2024). As technology advances and new communication streams evolve, more work is needed to understand how college students perceive the information they receive from these technologies and the impact of trust on the tools' efficacy.

Our work adds to a burgeoning literature around how clearer communications about course expectations can improve student performance. The closest study to ours is by Carrell and Kurlaender (2023). In this paper, the authors tested the effect of emails from course faculty providing students with feedback on their grades and encouraging them to access supplemental supports. Their pilot implementation (N professor=1, N students=69) targeted students who had not submitted the first course assignment (and were therefore starting the semester behind) and found an 8-percentage point increase in students' final course grade. Their scale-up (N=34 professors, 4,000 students) targeted all students enrolled in the participating courses and found precise null overall effects of the intervention on final course grades, though significant heterogenous treatment effects. For example, the email communication increased the likelihood freshmen earned an A or B by nine percentage points. The authors note that context matters for anticipating the potential efficacy of outreach.

Despite our consistent effects across subjects and faculty in these analyses, we hypothesize that the marginal benefit of the academic chatbot may also vary by course contexts. We targeted large, online, asynchronous courses precisely because they were settings where many students struggle to complete the course and earn high grades, and the effects of the academic chatbot may be smaller to null in courses with lower DFW rates (e.g., small seminar courses). The tool may also be less effective in courses where professors are already engaging in high-touch reminders and communication with students, though the professors teaching the focal courses in this RCT did engage in some personalized email communications with students. We also note that while many of the key components of the intervention – breaking down large assignments into manageable tasks, providing customized information about student performance to date, and opening a line of communication between the students and instructional team – translate across college subjects and courses, some features such as #quizme or specific questions about course content may be more difficult to scale. This tool is well suited to helping students through the administrative tasks of college courses and may be less effective in courses where the primary barrier to success is the difficulty of the course content or students' preparation for the course (e.g., computational skills). Future work on which we are embarking will explore the implementation and effectiveness of the academic chatbot across other subjects and within different course structures (e.g., in-person or smaller courses).

REFERENCES

Allen, J., & Robbins, S. B. (2008). Prediction of college major persistence based on vocational interests, academic preparation, and first-year academic performance. Research in Higher Education, 49, 62–79.

Bayer, A., Bruich, G., Chetty, R. & Housiaux, A. (2020) Expanding and diversifying the pool of undergraduates who study economics: Insights from a new introductory course at Harvard, The Journal of Economic Education, 51:3-4, 364-379, DOI: 10.1080/00220485.2020.1804511

Bayer, A. & Rouse, C. (2016). Diversity in the Economics Profession: A New Attack on an Old Problem. *Journal of Economic Perspectives*, 30 (4): 221-42.

Belley, P., & Lochner, L. (2007). The changing role of family income and ability in determining educational attainment. *Journal of Human Capital, 1*(1)

Bettinger, E., & Baker, R. (2014). The effects of student coaching: An evaluation of a randomized experiment in student advising. *Educational Evaluation and Policy Analysis, 36*(1), 3-19.

Bettinger, E., Doss, C., Loeb, S., Rogers, A., & Taylor, E. (2017). The effects of class size in online college courses: Experimental evidence. *Economics of Education Review, 5*8, 68-85.

Bettinger, E. P., Fox, L., Loeb, S., & Taylor, E. S. (2017). Virtual Classrooms: How Online College Courses Affect Student Success. *American Economic Review*, 107 (9): 2855-75.

Bick, A., Blandin, A., & Deming, D. (2024). The rapid adoption of generative AI. (No. w32966). National Bureau of Economic Research.

Bird, K., Castleman, B. L., & Lohner, G. (2022). Negative impacts from the shift to online learning during the COVID-19 crisis: Evidence from a statewide community college system. *AERA Open, 8*.

Boucher, K., Murphy, M., Bartell, D., Smail, J., Logel, C., & Danek, J. (2021) Centering the Student Experience: What Faculty and Institutions Can Do to Advance Equity, *Change: The Magazine of Higher Learning*, *53*:6, 42-50, DOI: 10.1080/00091383.2021.1987804

Carrell, S. E. & Kurlaender, M. (2020). My professor cares: Experimental evidence on the role of faculty engagement (No. w27312). National Bureau of Economic Research.

Castleman, B. L., & Long, B. T. (2016). Looking beyond enrollment: The causal effect of need-based grants on college access, persistence, and graduation. *Journal of Labor Economics*, *34*(4).

Castleman, B. L., & Page, L. C. (2015). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic Behavior & Organization*, *115*, 144-160.

Castleman, B. L., & Page, L. C. (2016). Freshman year financial aid nudges: An experiment to increase FAFSA renewal and college persistence. *Journal of Human Resources*, *51*(2), 389-415.

Castleman, B. L., & Page, L. C. (2017). Parental influences on postsecondary decision making: Evidence from a text messaging experiment. *Educational Evaluation and Policy Analysis*, *39*(2), 361-377.

Castleman, B. L., Page, L. C., & Schooley, K. (2014). The forgotten summer: Does the offer of college counseling after high school mitigate summer melt among college-intending, low-income high school graduates? *Journal of Policy Analysis and Management*, *33*(2), 320-344.

Coffman, K. B., Collis, M., & Kulkarni, L. (2024). Stereotypes and Belief Updating. *Journal of the European Economic Association 22*(3): 1011–1054.

Cohen, E., Huo, H., Guyot, K., Gaffney, C. & Christopher, E. (2024). A first look at the 2021 postsecondary enrollment, completion ,and financial aid outcomes of fall 2009 ninth-graders. NCES 2024-022. *National Center for Education Statistics*.

College Scorecard (n.d.). Georgia State University Profile. Retrieved from https://collegescorecard.ed.gov/.

Clotfelter, C. T., Hemelt, S. W., & Ladd, H. F. (2018). Multifaceted aid for low-income students and college outcomes: Evidence from North Carolina. *Economic Inquiry*, *56*(1), 278-303.

De Giorgi, G., Pellizzari, M, & Woolston, W. G. (2009). Class size and class heterogeneity. IZA Discussion Papers 4443, Institute of Labor Economics (IZA).

Dynan, K. E., & Rouse, C. E. (1997). The underrepresentation of women in economics: A study of undergraduate economics students. *The Journal of Economic Education*, *28*(4), 350-368.

Giani, M. S., & Martin, A. (2021). Mobilizing developmental education: The causal effect of mobile app courseware on the college outcomes of developmental education students. *Educational Evaluation and Policy Analysis, 43*(4), 668-687.

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology, 102*(4), 700–717. https://doi.org/10.1037/a0026659

Goulas, S. (2024). Twelve facts about the economics of education. The Hamilton Project, The Brookings Institution. Retrieved from https://www.brookings.edu/wp-content/uploads/2034/06/20240627_THP_EducationFacts_PDF.pdf

Holzer, H., & Baum, S., (2017). *Making College Work: Pathways to Success for Disadvantaged Students*. Washington, DC: Brookings Institution Press.

Jack, A. (2016). *The Privileged Poor: How Elite Colleges are Failing Disadvantaged Students*. Cambridge, MA: Harvard University Press.

Jurenka, I., …, Ibrahim, L. (2024). Towards responsible development of generative AI for education: An evaluation-driven approach. Google Working Paper. Retrieved from https://storage.googleapis.com/deepmind-media/LearnLM/LearnLM_paper.pdf

Kena, G., Aud, S., Johnson, F., Wang, X., Zhang, J., Rathbun, A., ... & Kristapovich, P. (2014). The Condition of Education 2014. NCES 2014-083. *National Center for Education Statistics*.

Kofoed, M., Gebhart, L., Gilmore, D., & Moschitto, R. (2021). Zooming to class?: Experimental evidence on college students' online learning during COVID-19. IZA Discussion Papers 14356, Institute of Labor Economics (IZA).

Lareau, A. (2003). *Unequal childhoods: Race, class, and family life.* (Second, Ed.) Oakland, CA: University of California Press.

Linkow, T., Miller, H., Parsad, A., Price, C., & Martinez, A. (2021). *Study of College Transition Messaging in GEAR UP: Impacts on Enrolling and Staying in College*. Institute of Education Sciences

Long, B. T., & Mabel, Z. (2012). Barriers to college success: Income disparities in progress to completion. Unpublished manuscript. Harvard University.

Ma, J., Pender, M, & Welch, M. (2019). Education Pays: 2019. Retrieved from: https://research.collegeboard.org/media/pdf/education-pays-2019-full-report.pdf

Marsicano, C., Felten, K., Toledo, L., & Buitendorp, M. (2020). Tracking campus responses to the COVID-19 pandemic. *APSA Preprints*. doi: 10.33774/apsa-2020-3wvrl.

Mendes, W. B., Gray, H. M., Mendoza-Denton, R., Major, B., & Epel, E. S. (2007). Why egalitarianism might be good for your health: physiological thriving during stressful intergroup encounters. *Psychological science*, *18*(11), 991–998. https://doi.org/10.1111/j.1467-9280.2007.02014.

NCES (2019). U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS) Fall Enrollment component, Spring 2013 through Spring 2019.

Oreopoulos, P., & Petronijevic, U. (2019). *The remarkable unresponsiveness of college students to nudging and what we can learn from it* (No. w26059). National Bureau of Economic Research.

Ortagus, J., Tanner, M., & McFarlin, I. (2020). Can re-enrollment campaigns help dropouts return to college? Evidence from Florida community colleges. *Educational Evaluation and Policy Analysis, 43*(1), 154-171.

Page, L. C., Castleman, B. L., & Meyer, K. (2020). Customized nudging to improve FAFSA completion and income verification. *Educational Evaluation and Policy Analysis*, *42*(1), 3-21.

Page, L. C., & Gehlbach, H. (2017). How an artificially intelligent virtual assistant helps students navigate the road to college. *AERA Open*, *3*(4).

Page, L. C., Kehoe, S. S., Castleman, B. L., & Sahadewo, G. A. (2019). More than dollars for scholars: The impact of the Dell Scholars Program on college access, persistence and degree attainment. *Journal of Human Resources, 54*(3), 683-725.

Page, L. C., Meyer, K., Lee, J., & Gehlbach, H. (2023). Conditions under which college students can be responsive to nudging. EdWorkingPapers Working Paper No. 20-242. Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/vjfs-kv29

Paunesku, D., & Farrington, C. A. (2020). Measure learning environments, not just students, to support learning and development. *Teachers College Record*, *122*(14), 1-26.

Pugatch, T., & Wilson, N. (2024). Nudging Demand for Academic Support Services: Experimental and Structural Evidence from Higher Education. *Journal of Human Resources*, *59*(5), 1637-1682

Scrivener, S., Weiss, M. J., Ratledge, A., Rudd, T., Sommo, C., Fresques, H. (2015). Doubling graduation rates: Three-year effects of CUNY's Accelerated Study in Associate Programs (ASAP) for developmental education students. MDRC evaluation report. Retrieved from: https://www.mdrc.org/sites/default/files/doubling_graduation_rates_fr.pdf

Scuello, M., & Strumbos, D. (2024). Evaluation of Accelerate, complete, engage (ACE) at CUNY John Jay College of Criminal Justice: Final Report (ASAP/ACE (CUNY John Jay)). https://www.cuny.edu/wp-content/uploads/sites/4/page-assets/about/administration/offices/student-success-initiatives/asap/about/ace/300414_CUNY_March_2024_ACE_Final_Report_m1-1.pdf

Simon, H. A. (1982). *Models of bounded rationality.* Cambridge, MA: MIT Press

Smith, E. N. (2020). Expansive care theory: Can messages of inspiring expectations and broad regard promote students' identity safety and academic success? [Doctoral dissertation, Stanford University] Available from ProQuest Dissertations & Theses Global.

Smith, B. O., White, D. R., Kuzyk, P. C., & Tierney, J. E. (2018) Improved grade outcomes with an e-mailed "grade nudge", *The Journal of Economic Education, 49*:1, 1-7, DOI: 10.1080/00220485.2017.1397570

Snyder, T., & Dillow, S. (2015). Digest of Education Statistics 2013. Retrieved from https://nces.ed.gov/pubs2015/2015011.pdf.

Sommo, C., Slaughter, A., Saunier, C., Scrivener, S., & Warner, K. (2023). *Varying Levels of Success.* MDRC.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness.* London: Penguin Books.

Toppo, G. (2024, August 7). AI 'companions' are patient, funny, upbeat – and probably rewiring kids' brains. *The 74 Million*.

Walton, G., & Cohen, G. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology, 92*(1), 82-96.

Wang, R., Ribeiro, A., Robinson, C., Loeb, S., & Demszky, D. (2024). Tutor CoPilot: A human-AI approach for scaling real-time expertise. EdWorkingPapers Working Paper No. 24-1054. Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/81nh-8262

Wanless, S. B. (2016). The role of psychological safety in human development. *Research in Human Development*, *13*(1), 6-14.

Weiss, M. J., Ratledge, A., Sommo, C., & Gupta, H. (2019). supporting community college students from start to degree completion: Long-term evidence from a randomized trial of CUNY's ASAP. *American Economic Journal: Applied Economics*, *11*(3), 253-97.

Wells, G. (2023, August 24). Gen Z-ers are computer whizzes. Just don't ask them to type. *The Wall Street Journal*.

Yellen, J. (2019, September 24). Former Fed chair Janet Yellen on gender and racial diversity of the federal government's economists. The Brookings Institution forum on "The gender and racial diversity of the federal government's economics."

TABLES

Table 1: Analytic Sample and Randomization Balance

	Panel A: Pooled across courses			Panel B: Government			Panel C: Microeconomics		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Control mean	Treatment effect	N	Control mean	Treatment effect	N	Control mean	Treatment effect	N
Female	0.54	0.02	2483	0.56	0.03	1568	0.51	0.00	915
		(0.020)			(0.025)			(0.033)	
Asian	0.22	0.00	2483	0.23	0.00	1568	0.20	0.01	915
		(0.017)			(0.021)			(0.026)	
Black	0.46	0.03	2483	0.43	0.04	1568	0.51	0.01	915
		(0.020)			(0.025)			(0.033)	
White	0.22	-0.03	2483	0.23	-0.03	1568	0.20	-0.02	915
		(0.016)			(0.021)			(0.025)	
Hispanic	0.14	-0.01	2483	0.15	-0.01	1568	0.12	0.00	915
		(0.014)			(0.018)			(0.021)	
First Generation	0.22	0.02	2483	0.24	0.02	1568	0.20	0.01	915
		(0.017)			(0.022)			(0.027)	
Pell Eligible	0.57	0.00	2483	0.58	-0.02	1568	0.55	0.02	915
		(0.020)			(0.025)			(0.033)	
Course Re-takers	0.11	0.01	2483	0.08	0.01	1568	0.15	0.01	915
		(0.012)			(0.014)			(0.024)	
Freshman	0.52	-0.01	2483	0.63	0.01	1568	0.33	-0.04	915
		(0.017)			(0.022)			(0.028)	
Upperclassman	0.40	0.00	2483	0.30	-0.01	1568	0.55	0.03	915
		(0.018)			(0.021)			(0.032)	
Transfer	0.08	0.01	2483	0.07	0.00	1568	0.12	0.02	915
		(0.011)			(0.013)			(0.021)	
High School GPA	3.47	0.02	2043	3.52	0.01	1393	3.37	0.05	650
		(0.017)			(0.020)			(0.033)	
Joint F-Test		0.5571			0.4746			0.7442	

Notes: Robust standard errors in parentheses. Includes randomization blocks. Models pooling across subjects include subject fixed effects. High school GPA reported here excludes missing cases.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 2: Intent-to-treat effect of course chatbot on final grades

	Panel A: Pooled across courses					Panel B: Government					Panel C: Microeconomics				
	(1)	(2)		(3)		(4)	(5)	(6)		(7)	(8)		(9)	
	Control mean	Treatment effect		Treatment effect		Control mean	Treatment effect	Treatment effect		Control mean	Treatment effect		Treatment effect	
Final Grade	71.07	1.58		1.80		71.88	0.81	1.24		69.66	2.90		3.07	
		(1.283)		(1.226)			(1.539)	(1.463)			(2.279)		(2.193)	
Earned A	0.36	0.04	+	0.04	*	0.31	0.04	0.04	+	0.44	0.03		0.04	
		(0.019)		(0.018)			(0.024)	(0.022)			(0.032)		(0.031)	
Earned B or higher	0.61	0.04	*	0.04	**	0.61	0.04	0.04	*	0.62	0.05		0.05	+
		(0.019)		(0.018)			(0.024)	(0.022)			(0.031)		(0.030)	
Earned C or higher	0.73	0.02		0.02		0.75	0.01	0.02		0.71	0.04		0.04	
		(0.017)		(0.017)			(0.022)	(0.021)			(0.029)		(0.028)	
D or F	0.15	-0.01		-0.02		0.15	-0.02	-0.02		0.15	0.00		0.00	
		(0.014)		(0.013)			(0.017)	(0.017)			(0.023)		(0.023)	
Withdrew	0.04	0.00		0.00		0.03	0.00	0.00		0.07	0.00		-0.01	
		(0.008)		(0.008)			(0.008)	(0.008)			(0.016)		(0.016)	
DFW	0.19	-0.02		-0.02		0.18	-0.03	-0.03	+	0.22	-0.01		-0.01	
		(0.015)		(0.015)			(0.019)	(0.018)			(0.027)		(0.026)	
Dropped	0.07	0.00		0.00		0.07	0.01	0.01		0.08	-0.03	+	-0.03	+
		(0.010)		(0.010)			(0.013)	(0.014)			(0.016)		(0.016)	
Covariates included				X				X					X	
N students		2483		2483			1568	1568			915		915	

Notes: Robust standard errors in parentheses. Includes randomization blocks. Models pooling across subjects include subject fixed effects. "DFW" stands for earning a D or F in the course or withdrawing from the course. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA (imputed as zero for missing cases, with a covariate flag for having a missing GPA).
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 3: Treatment effect of chatbot on final grades, by gender, Microeconomics

	Panel A: Women			Panel B: Men		
	(1)	(2)		(3)	(4)	(5)
	Control Average	Treatment Effect		Control Average	Treatment Effect	Test of Equality
Final Grade	68.86	7.24	**	70.48	-0.34	0.071
		(3.022)			(3.182)	
Earned A	0.43	0.07		0.46	0.00	0.225
		(0.044)			(0.044)	
Earned B or higher	0.60	0.11	**	0.64	0.01	0.056
		(0.042)			(0.042)	
Earned C or higher	0.71	0.10	**	0.71	-0.01	0.034
		(0.039)			(0.041)	
D or F	0.14	-0.03		0.16	0.02	0.259
		(0.031)			(0.034)	
Withdrew	0.07	-0.01		0.07	0.00	0.803
		(0.021)			(0.025)	
DFW	0.21	-0.04		0.22	0.02	0.249
		(0.036)			(0.038)	
Dropped	0.09	-0.06	**	0.07	0.00	0.071
		(0.023)			(0.025)	
Covariates included		X			X	
N students		463			452	915

Notes: Robust standard errors in parentheses. Includes randomization blocks. "DFW" stands for earning a D or F in the course or withdrawing from the course. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. Test of equality evaluates equality of the treatment effect coefficient from separate regressions.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 4: Completion of and Performance on Government Assignments

	(1) Control Mean	(2) Treatment Effect	(3) Treatment Effect
Final grade	71.88	0.81	1.24
		(1.539)	(1.463)
Reading score	78.31	0.39	0.74
		(1.679)	(1.623)
Completed Exam 1	0.86	0.00	0.00
		(0.018)	(0.018)
Performance on Exam 1	65.75	-0.31	0.23
		(1.505)	(1.425)
Complete Exam 2	0.83	0.01	0.01
		(0.019)	(0.019)
Performance on Exam 2	60.22	1.29	1.80
		(1.525)	(1.446)
Completed Exam 3	0.82	0.00	0.01
		(0.019)	(0.019)
Performance on Exam 3	62.62	0.57	1.09
		(1.591)	(1.516)
N students		1568	1568
Completed Exam 4	0.81	0.02	0.02
		(0.034)	(0.033)
Performance on Exam 4	58.14	1.30	1.75
		(2.617)	(2.483)
N students		509	509
Completed Field Trip	0.79	0.03	0.03
		(0.025)	(0.024)
Grade on Field Trip	83.20	3.08	3.05
		(2.768)	(2.644)
Reading minutes	578.85	2.92	1.60
		(21.018)	(20.780)
N students		990	990
Covariates included			X

Notes: Robust standard errors in parentheses. Includes randomization blocks. Course assignments changed across intervention terms; exam 4 was only administered the first intervention term and the field trip assignment was only required the first two intervention terms. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. +p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 5: Completion of and Performance on Microeconomics Assignments

		(1)	(2)	(3)	
		Control Mean	Treatment Effect	Treatment Effect	
Final course grade		69.66	2.90	3.31	
			(2.279)	(2.175)	
Participation (10%)	Grade on Pre/Post Quizzes	73.50	1.81	2.21	
			(2.485)	(2.393)	
	Completed Discussion Posts	0.61	0.00	0.00	
			(0.032)	(0.031)	
	Grade in Discussion Posts	65.20	1.16	1.54	
			(2.637)	(2.529)	
Practice Interactive Tools (15%)	Completed	0.67	0.03	0.04	
			(0.030)	(0.029)	
	Grade	77.05	3.37	3.86	
			(2.487)	(2.407)	
Assessment Interactive Tools (25%)	Completed	0.62	0.02	0.03	
			(0.032)	(0.031)	
	Grade	69.97	2.56	3.04	
			(2.440)	(2.333)	
Practice Quizzes (15%)	Completed	0.74	0.04	0.05	+
			(0.028)	(0.027)	
	Grade	70.49	3.45	3.79	+
			(2.234)	(2.138)	
Assessment Quizzes (35%)	Completed	0.60	0.01	0.01	
			(0.032)	(0.030)	
	Grade	64.10	3.08	3.47	+
			(2.184)	(2.076)	
N students			915	915	
Covariates included				X	

Notes: Robust standard errors in parentheses. Includes randomization blocks. Percentages listed next to assignment components reference the weight each assignment received in final grade calculations. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. +p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 6: Treatment effect on take-up of supplemental instruction

	(1)	(2)		(3)	
	Control Mean	Treatment Effect		Treatment Effect	
Pooled: Used SI	0.05	0.02	*	0.02	*
		(0.010)		(0.010)	
N students		2483		2483	
American Government: Used SI	0.07	0.03	+	0.02	
		(0.014)		(0.014)	
N students		1568		1568	
Microeconomics: Used SI	0.02	0.02		0.02	
		(0.012)		(0.012)	
N students		915		915	
Covariates included				X	

Notes: Robust standard errors in parentheses. Includes randomization blocks. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 7: Treatment effect on student attitudes, Government

	(1)	(2)	(3)	(4)
	Control Mean	Treatment	Treatment	N
Completed survey	0.67	0.01	0.01	1568
		(0.023)	(0.022)	
Organizational Support (1-6)	4.45	0.04	0.03	1063
		(0.073)	(0.072)	
Self-Efficacy (1-6)	4.67	0.00	0.00	1063
		(0.073)	(0.072)	
Adaptive Student Attributions (1-5)	3.81	0.01	0.00	1063
		(0.047)	(0.047)	
Perception of Achievable Challenge (1-6)	3.78	0.00	-0.02	1061
		(0.048)	(0.048)	
Sense of Social Belonging (1-6)	4.20	-0.05	-0.07	1056
		(0.053)	(0.053)	
Trust and Fairness (1-6)	5.01	0.04	0.04	1055
		(0.053)	(0.052)	
Meaningful Work (1-6)	4.73	0.03	0.02	1065
		(0.057)	(0.056)	
Level of Nervousness with Instructor (1-5)	2.52	0.07	0.08	1061
		(0.067)	(0.066)	
Broad Regard (1-6)	4.03	0.04	0.05	1061
		(0.061)	(0.062)	
Covariates included			X	

Notes: Robust standard errors in parentheses. Includes randomization blocks. Item scale in parentheses next to index. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. Sample size reported separately for each construct measured.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Table 8: Treatment effects on non-course academic outcomes

		Pooled				American Government					Microeconomics			
		Control Mean	Treatment	Treatment	N	Control Mean	Treatment	Treatment		N	Control Mean	Treatment	Treatment	N
Overall semester performance	Term GPA including focal course	2.57	0.05	0.06	2483	2.60	0.08	0.11	+	1568	2.52	-0.02	-0.01	915
			(0.053)	(0.048)			(0.066)	(0.059)				(0.088)	(0.083)	
Performance in other courses	Earned credits outside of focal course	0.90	0.00	0.00	2483	0.91	-0.01	-0.01		1568	0.88	0.02	0.02	915
			(0.012)	(0.012)			(0.015)	(0.015)				(0.021)	(0.021)	
	Term GPA in other courses (zero if missing)	2.43	0.01	0.02	2483	2.50	0.05	0.07		1568	2.31	-0.05	-0.04	915
			(0.057)	(0.053)			(0.071)	(0.065)				(0.094)	(0.089)	
	Term credits in other courses (zero if missing)	7.65	0.09	0.13	2483	8.04	0.18	0.23		1568	6.98	-0.06	0.01	915
			(0.180)	(0.165)			(0.225)	(0.206)				(0.296)	(0.276)	
	Term GPA in other courses (zero only if withdrew from other courses)	2.67	0.01	0.03	2263	2.72	0.08	0.09		1437	2.58	-0.09	-0.06	826
			(0.053)	(0.049)			(0.066)	(0.059)				(0.089)	(0.085)	
	Term credits in other courses (zero only if withdrew from other courses)	8.40	0.11	0.17	2263	8.74	0.28	0.32	+	1437	7.79	-0.15	-0.04	826
			(0.167)	(0.151)			(0.204)	(0.183)				(0.280)	(0.265)	
Persistence in subject	Took a course in department next term	0.13	0.00	0.00	2483	0.07	-0.01	-0.01		1568	0.24	0.01	0.01	915
			(0.013)	(0.013)			(0.013)	(0.013)				(0.028)	(0.028)	
Covariates included				X				X					X	

Notes: Robust standard errors in parentheses. Includes randomization blocks. Models pooling across subjects include subject fixed effects. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the

student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. Whether student has a spillover measure is an indicator for whether the student (1) completed the intervention course and (2) completed at least one other course that semester. Term GPA (measured on a 4.0 scale) and term hours (with each GSU course bearing about 3 credit hours) calculated only for students with a spillover measure.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

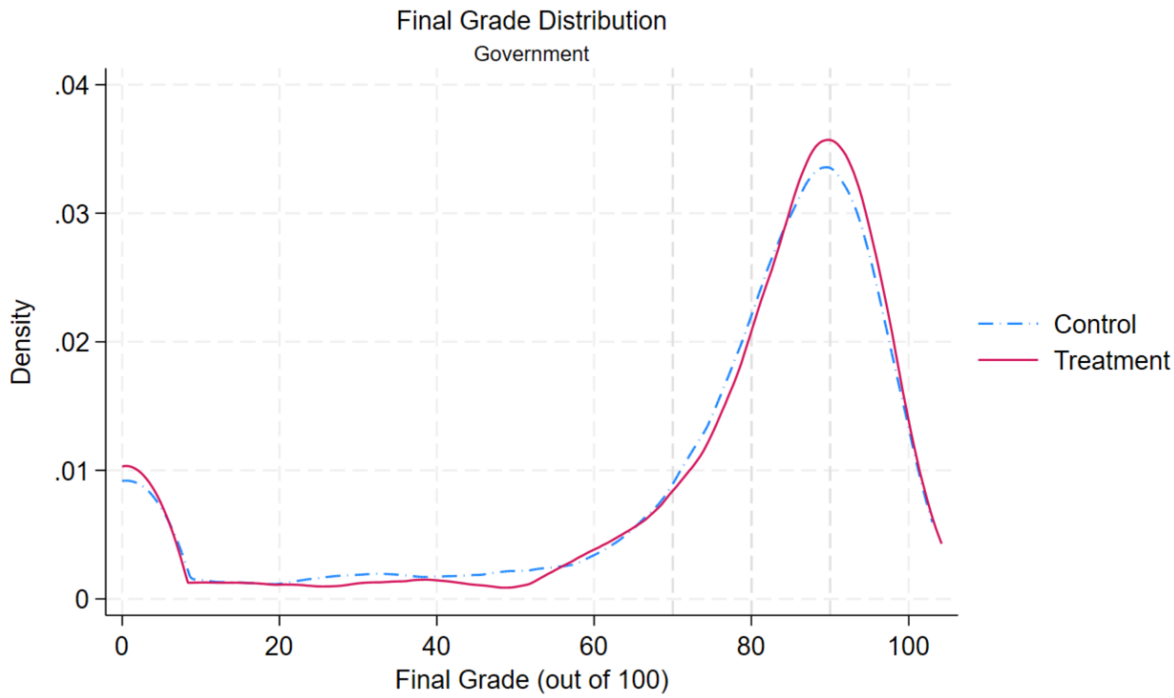
Table 9: Chatbot engagement summary, treated students

	(1)	(2)	(3)	(4)	(5)
	Ever received	Total received	Ever opt out	Ever Reply	Total Replies
American Government	0.99	46.01	0.04	0.54	4.63
	[0.080]	[23.636]	[0.206]	[0.499]	[9.172]
Microeconomics	1.00	44.40	0.03	0.43	1.07
	[0.066]	[12.862]	[0.165]	[0.496]	[2.087]

Notes: Summarizes chatbot engagement among treated students. Standard deviations in brackets.

FIGURES

Figure 1. Final grade density



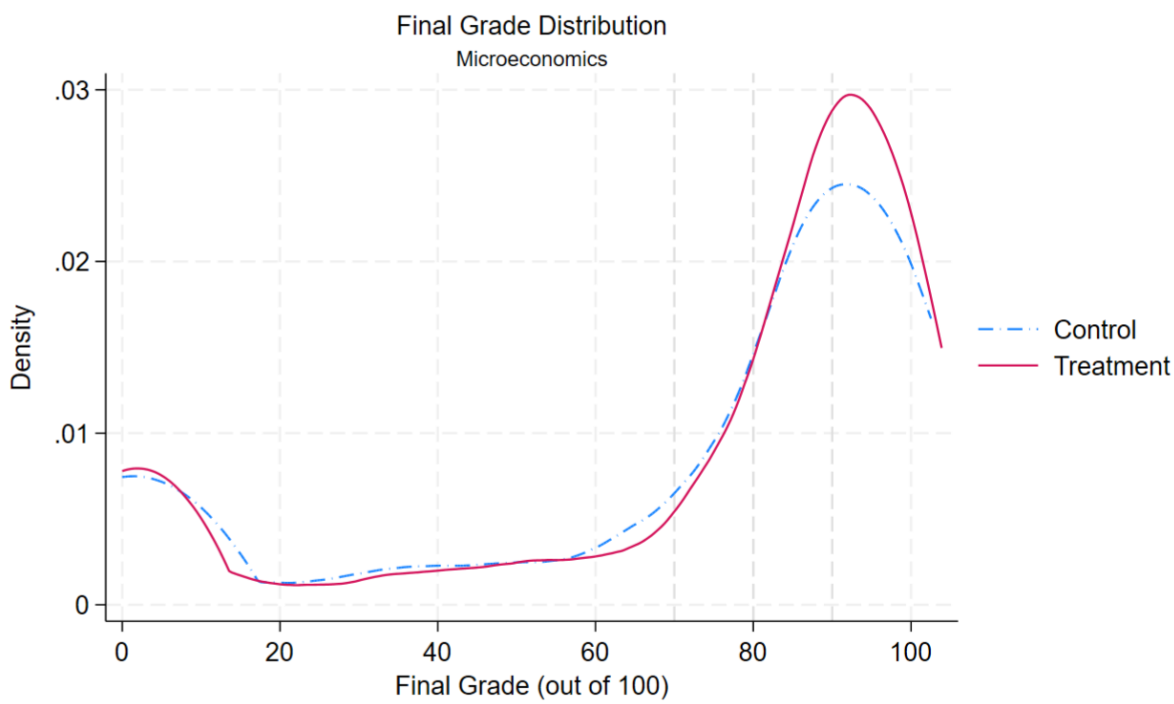


Figure 2a. Heterogeneous treatment effect of course chatbot, pooled across subjects

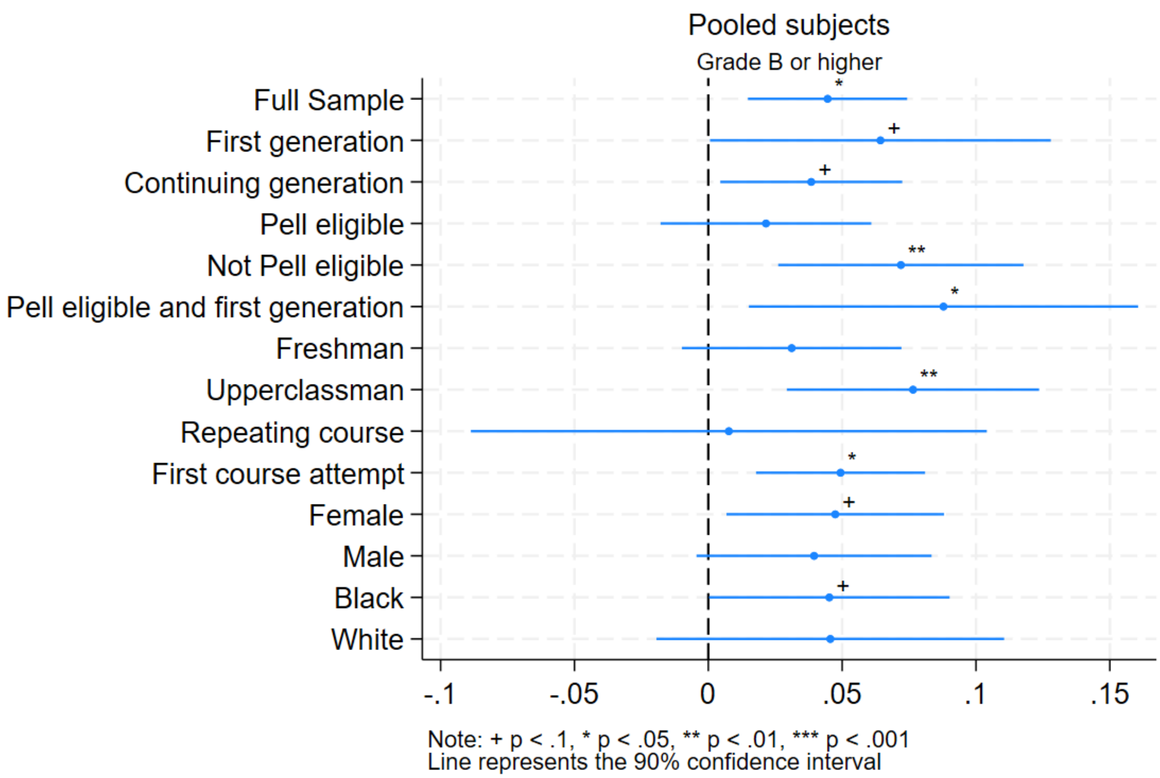


Figure 2b. Heterogeneous treatment effect of course chatbot, Government

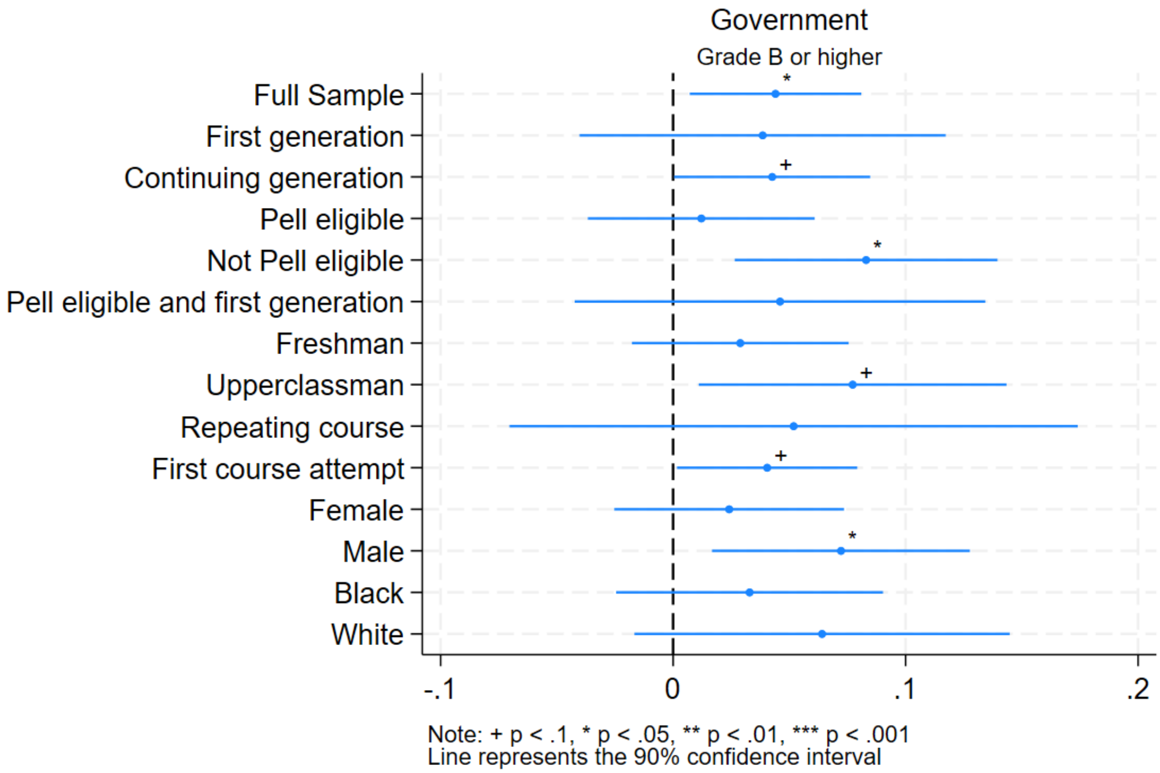


Figure 2c. Heterogeneous treatment effect of course chatbot, Microeconomics

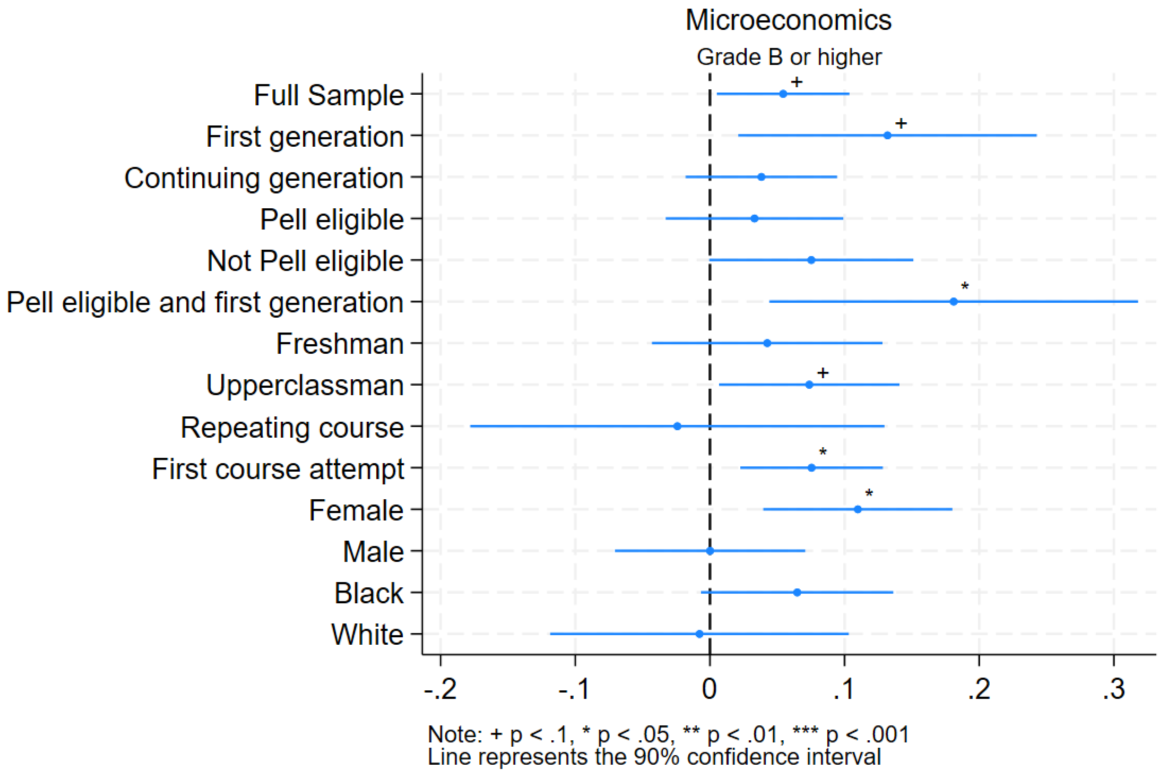


Figure 3. Heterogeneous treatment effect of course chatbot on supplemental instruction attendance (SI), full sample

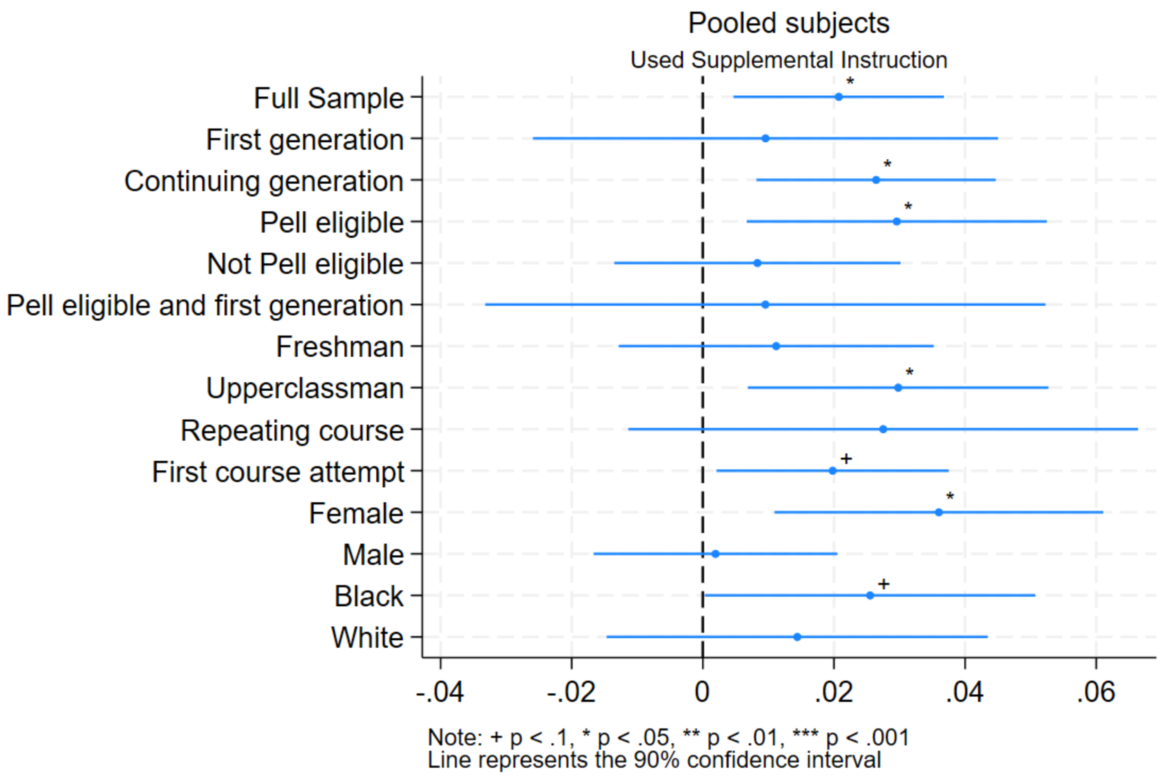


Figure 4. Heterogeneous treatment effect of course chatbot on semester GPA, by course

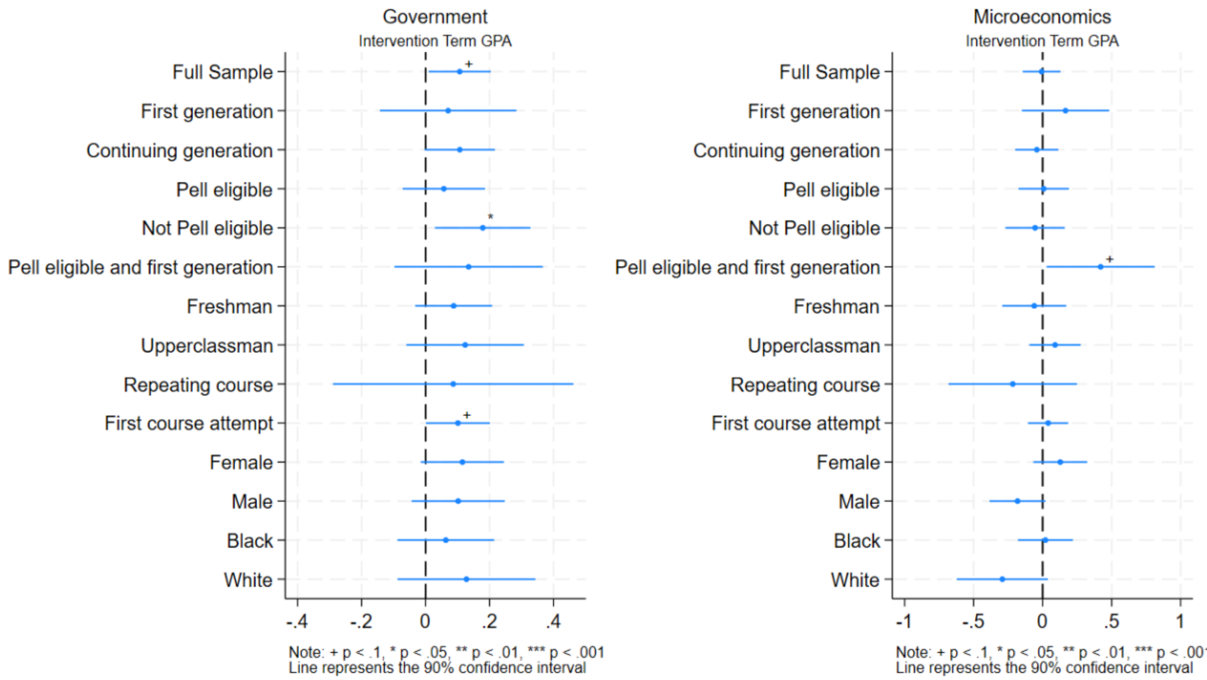
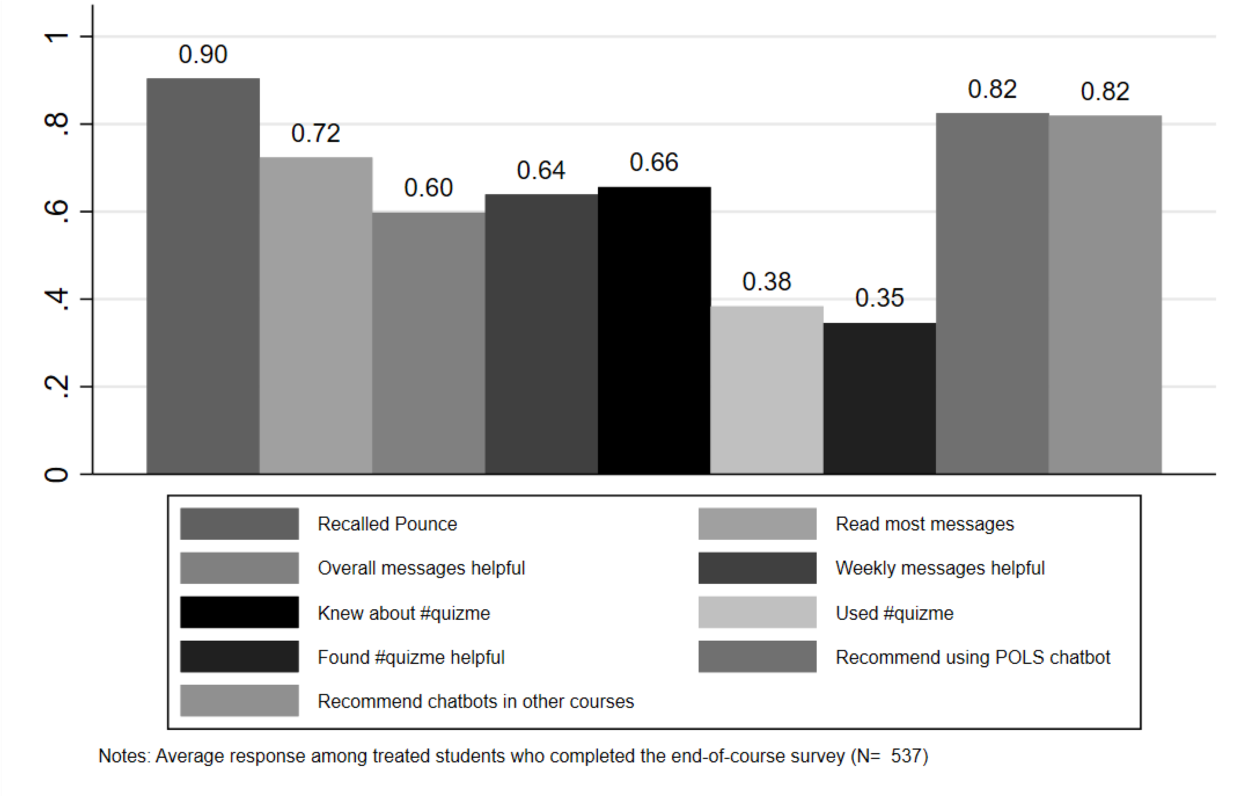


Figure 5. Student survey responses, Government



APPENDIX TABLES

Appendix Table 1: Analytic Sample and Randomization Balance, Government

	Fall 2021		Spring 2022		Fall 2022		Pooled	
	Control	Treatment difference	Control	Treatment difference	Control	Treatment difference	Control	Treatment difference
Female	0.62	0.01	0.60	0.06	0.47	0.02	0.56	0.03
		(0.043)		(0.044)		(0.042)		(0.025)
Asian	0.19	-0.01	0.27	-0.01	0.24	0.01	0.23	0.00
		(0.034)		(0.040)		(0.036)		(0.021)
Black	0.46	0.06	0.43	0.04	0.40	0.03	0.43	0.04
		(0.044)		(0.045)		(0.041)		(0.025)
White	0.27	-0.05	0.20	-0.01	0.23	-0.02	0.23	-0.03
		(0.038)		(0.036)		(0.035)		(0.021)
Hispanic	0.15	-0.01	0.15	0.01	0.15	-0.03	0.15	-0.01
		(0.031)		(0.033)		(0.029)		(0.018)
First Generation	0.24	0.01	0.24	0.04	0.23	0.02	0.24	0.02
		(0.039)		(0.040)		(0.035)		(0.022)
Pell Eligible	0.63	-0.04	0.59	0.01	0.53	-0.02	0.58	-0.02
		(0.043)		(0.045)		(0.042)		(0.025)
Course Re-takers	0.07	0.01	0.13	0.02	0.07	0.00	0.08	0.01
		(0.023)		(0.031)		(0.019)		(0.014)
Freshman	0.43	0.05	0.65	0.00	0.78	-0.01	0.63	0.01
		(0.043)		(0.043)		(0.028)		(0.022)
Upperclassman	0.49	-0.04	0.27	-0.01	0.17	0.00	0.30	-0.01
		(0.043)		(0.040)		(0.027)		(0.021)
Transfer	0.07	0.00	0.08	0.01	0.05	0.01	0.07	0.00
		(0.023)		(0.025)		(0.018)		(0.013)
High School GPA	3.45	0.02	3.46	0.05	3.61	-0.02	3.52	0.01
		(0.035)		(0.036)		(0.032)		(0.020)
Joint F-Test		0.878		0.743		0.934		0.475
N students		509		481		578		1568

Notes: Robust standard errors in parentheses. Includes randomization blocks. High school GPA reported here excludes missing cases.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Appendix Table 2: Analytic Sample and Randomization Balance, Microeconomics

	Fall 2022		Spring 2023		Pooled	
	Control	Treatment difference	Control	Treatment difference	Control	Treatment difference
Female	0.54	-0.02	0.47	0.02	0.51	0.00
		(0.048)		(0.045)		(0.033)
Asian	0.20	-0.01	0.19	0.02	0.20	0.01
		(0.038)		(0.036)		(0.026)
Black	0.49	0.04	0.54	-0.02	0.51	0.01
		(0.048)		(0.045)		(0.033)
White	0.19	-0.03	0.20	-0.02	0.20	-0.02
		(0.037)		(0.035)		(0.025)
Hispanic	0.13	0.01	0.11	0.00	0.12	0.00
		(0.033)		(0.028)		(0.021)
First Generation	0.20	0.03	0.21	0.00	0.20	0.01
		(0.040)		(0.037)		(0.027)
Pell Eligible	0.57	-0.02	0.54	0.05	0.55	0.02
		(0.048)		(0.045)		(0.033)
Course Re-takers	0.14	-0.01	0.16	0.02	0.15	0.01
		(0.034)		(0.034)		(0.024)
Freshman	0.13	-0.03	0.50	-0.06	0.33	-0.04
		(0.031)		(0.045)		(0.028)
Upperclassman	0.70	0.00	0.42	0.05	0.55	0.03
		(0.044)		(0.045)		(0.032)
Transfer	0.16	0.03	0.07	0.01	0.12	0.02
		(0.037)		(0.024)		(0.021)
High School GPA	3.40	0.03	3.36	0.06	3.37	0.05
		(0.053)		(0.042)		(0.033)
Joint F-Test		0.9918		0.7759		0.7442
N students		426		489		915

Notes: Robust standard errors in parentheses. Includes randomization blocks. High school GPA means reported here include zeros for missing cases.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Appendix Table 3: Main outcomes omitting students who dropped the course during the add/drop period

	Control	Treatment Effect		Treatment Effect	
Early drop	0.04	0.01		0.01	
		(0.008)		(0.008)	
N students		2483		2483	
Final Grade	73.62	2.09	+	2.24	*
		(1.193)		(1.129)	
Earned A	0.37	0.04	+	0.03	+
		(0.020)		(0.018)	
Earned B or higher	0.63	0.05	**	0.05	**
		(0.019)		(0.018)	
Earned C or higher	0.76	0.03		0.03	+
		(0.017)		(0.016)	
DFW	0.20	-0.02		-0.02	
		(0.016)		(0.015)	
DFW or drop	0.24	-0.03		-0.03	+
		(0.017)		(0.016)	
Withdrew	0.04	0.00		-0.01	
		(0.008)		(0.008)	
Dropped	0.04	-0.01		-0.01	
		(0.008)		(0.008)	
N students		2374		2374	
Covariates included				X	

Notes: Robust standard errors in parentheses. Includes randomization blocks. Models pooling across subjects include subject fixed effects. "DFW" stands for earning a D or F in the course or withdrawing from the course. Models including covariates control for sex, race, whether student applied for financial aid, Pell grant eligibility, whether student was a first-generation college student, whether student had taken the course prior to this term, whether the student had ever enrolled in a course using a chatbot, their year in school, and their high school GPA. Sample limited to students who remained enrolled in the course following the institution add/drop deadline.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

Appendix Table 4: Selection into end-of-course survey completion, Government

	Didn't complete survey	Completed Survey	Difference	
Treatment	0.493	0.505	0.013	
			(0.028)	
Female	0.554	0.582	0.041	
			(0.028)	
Asian	0.141	0.276	0.124	***
			(0.022)	
Black	0.552	0.398	-0.148	***
			(0.028)	
White	0.204	0.229	0.035	
			(0.023)	
Hispanic	0.117	0.155	0.039	*
			(0.019)	
First Generation	0.259	0.244	-0.016	
			(0.024)	
Pell Eligible	0.604	0.558	-0.038	
			(0.028)	
Course Re-takers	0.147	0.060	-0.100	***
			(0.018)	
Freshman	0.578	0.661	0.022	
			(0.025)	
Upperclassman	0.358	0.268	-0.033	
			(0.025)	
Transfer	0.063	0.071	0.010	
			(0.015)	
High School GPA	0.513	0.395	-0.132	***
			(0.028)	
N students	505	1063	1568	

Notes: Robust standard errors in parentheses for model reporting difference in characteristics among the survey completers and non-completers; mode includes randomization blocks.
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

APPENDIX B – Sample chatbot Messages

LAUNCH MESSAGE_08.23.2021

Department / Office	Political Science 1101
Purpose	Launching TA Pounce to students in POLS 1101 (group 1)
Target Population	180
Successful Contacts	176
Script	

Hi name_first! I'm the chatbot for American Government.
This term I'm working with Dr. Evans to help you stay on track. I'll send you course reminders and tips to succeed. You can text me questions anytime! So hit me up and I'll do my best to get you the answer.

Contacts without this profile information receive a backup text.

Pro-tip: Start each week with Dr. Evans' announcement.
👉 bit.ly/pols1101ann
If you don't want these messages, just text #PAUSE to stop (but I hope you'll give me a chance).

WEEK 1 GENERAL_08.24.2021

Department / Office	Political Science 1101
Purpose	Weekly reminder of upcoming due dates sent to all students
Target Population	178
Successful Contacts	174
Script	

WEEKLY DIGEST 🤖
Hi - each week I'll send you a reminder of upcoming due dates. Last year, almost all students found these weekly digest messages helpful. They help you have all the info before you make a plan to complete your coursework.
Use this link to access Dr. Evans' announcement
👉 bit.ly/F21pols1101ann

PRO-TIP: Download the Exam 1 study guide and fill out Ch. 1 this week as you read.
DUE THIS WEEK: You have 6 tasks (2hrs. total) due 🧩 WK1 Due Date.
1 Watch COURSE INTRO VIDEO
2 Read Syllabus
3 Take Syllabus Quiz
4 Watch AREA9 INTRO VIDEO
5 Read Chapter 1
6 Take Pre-Course Survey
👉 bit.ly/F21pols1101toc

WEEK 3 CUSTOMIZED DIGEST_ALL COMPLETE_09.07.2021

Department / Office	Political Science 1101
Purpose	Weekly reminder of upcoming due dates + personalized message to students who have completed all previously due graded requirements
Target Population	180
Successful Contacts	173
Script	

WEEKLY DIGEST 🤖
Hi name_first! I see you've completed all your reading so far this semester. Keep up the good work!
Reviewing the readings will help with Exam 1 next week. Supplemental Instruction (SI) can also help.

Contacts without this profile information receive a backup text.

Did you know that students who regularly attend SI score an average of one letter grade higher than students who don't? Come check it out!
👉 BITLY: SI

PRO-TIP: Download & fill out the Exam 1 study guide. Schedule time in your calendar to fill out Chs. 1&2 if you haven't already.
DUE THIS WEEK: You have 2 tasks (about 2hrs total) due WK3 Due Date.
1 Read Chapter 3
2 Complete "Activity: Know Thy Political Self?"
👉 BITLY: TOC

WEEK 3 CUSTOMIZED DIGEST_MISSING_09.07.2021

Department / Office	Political Science 1101
Purpose	Weekly reminder of upcoming due dates + personalized message to students who have at least missing graded requirement (<70%)
Target Population	24
Successful Contacts	23
Script	

WEEKLY DIGEST 🐻

Hi name_first! You seem to have a missing assignment. Check your iCollege email to see how to make it up. I'm here to help, so text me with any questions. Exam 1 is next week. Supplemental Instruction (SI) is a great way to prepare.

Contacts without this profile information receive a backup text.

Did you know that students who regularly attend SI score an average of one letter grade higher than students who don't? Can you find an hour this week to attend SI?

👉 BITLY: SI

PRO-TIP: Download & fill out the Exam 1 study guide. Schedule time in your calendar to fill out Chs. 1&2 if you haven't already.

DUE THIS WEEK: You have 2 tasks (about 2hrs total) due WK3 Due Date.

1 Read Chapter 3

2 Complete "Activity: Know Thy Political Self?"

👉 BITLY: TOC

WEEK 3 CUSTOMIZED DIGEST_WORK AHEAD_09.07.2021

Department / Office	Political Science 1101
Purpose	Weekly reminder of upcoming due dates + personalized message to students who have already completed all graded requirements due in course so far including the current week--students have worked ahead
Target Population	6
Successful Contacts	5
Script	

WEEKLY DIGEST 🤓
Hi name_first! I see you've already completed the assignments for this week. That's truly awesome—keep up the excellent work!
Exam 1 is next week. Supplemental Instruction (SI) is a great way to review the reading you've done.

Contacts without this profile information receive a backup text.

Did you know that students who regularly attend SI score an average of one letter grade higher than students who don't? Come check it out:
👉 BITLY: SI

PRO-TIP: Download & fill out the Exam 1 study guide this week if you haven't already. This is also a great time to complete the NCCHR assignment early for extra credit.
Have a great week!

LAUNCH #QUIZME/INTRO TYLER_09.10.2021

Department / Office	Political Science 1101
Purpose	Encouraging message to all students introducing Tyler (but not COMMAND #tyler) and introducing COMMAND #quizme
Target Population	239
Successful Contacts	235

Script

Howdy name_first! This is Tyler—the human behind the chatbot for American Gov. Many students have told me it can be hard to know if they have studied enough for a test. So I've set up a feature for you in this chatbot called: #quizme

Contacts without this profile information receive a backup text.

Text back the command #quizme (include the hashtag) anytime to start a short quiz with questions covering concepts on Exam 1 coming up in 5 days. After you take the quiz, hit me up with any questions. Teamwork makes the dream work. Let's be a team!

WEEK 5 CUSTOMIZED DIGEST_ MISSING EXAM 1_ 09.20.2021

Department / Office	Political Science 1101
Purpose	Week 5 message to students who did not complete Exam 1.
Target Population	10
Successful Contacts	10
Script	

WEEKLY DIGEST 🤖
Hi name_first! I see you did not take Exam 1. Follow this link to take action on making up the exam. Do this today! Dr. Evans will only allow make-ups for a few days.
👉 bit.ly/exam1makeup

Contacts without this profile information receive a backup text.

PRO-TIP: If you complete the NCCHR assignment this week, it could add up to 3 points to your final course grade.
DUE THIS WEEK: You have 2 tasks (about 2hrs. total) due WK5 Due Date.
1 Read Chapter 4
2 Take Check-in Survey I
👉 BITLY: TOC

ENCOURAGEMENT WK5_09.23.2021

Department / Office	Political Science 1101
Purpose	Encouragement message sent to all students addressing how students may be feeling overwhelmed at this point in the semester.
Target Population	225
Successful Contacts	215
Script	

Howdy name_first! Tyler here—the human behind the chatbot for American Gov. Students have told me they feel overwhelmed at this point in the semester. Especially after the first exam, it's totally normal to feel this way.

Contacts without this profile information receive a backup text.

It's also totally normal for this feeling to pass, so I encourage you to continue working hard. I'm here for you, too. Text in your questions any time and if the bot can't answer them, I will. I wish you the best of luck this semester. I'm rooting for you big time!

ENCOURAGEMENT INTERACTIVE WK10_10.28.2021

Department / Office	Political Science 1101
Purpose	Encouraging message sent to all student asking them to share how they felt about Exam 2.
Target Population	169
Successful Contacts	155
Response Rate	29%
Script	

Howdy name_first! Tyler here! Sooo.. how'd the exam go?! I know Exam 2 is typically the hardest one. How did you feel about it? REPLY 1/2/3
Contacts without this profile information receive a backup text.

[1]: Good! 😁
[2]: Meh... 😐
[3]: Not so good. 😒

1 Good! 😁

That's great to hear! So that we can better support our students, would you mind sharing what you found most helpful during studying or on the exam? I'm sure your peers would love to hear any advice you have to share. We won't include your name!

2 Meh... 😐

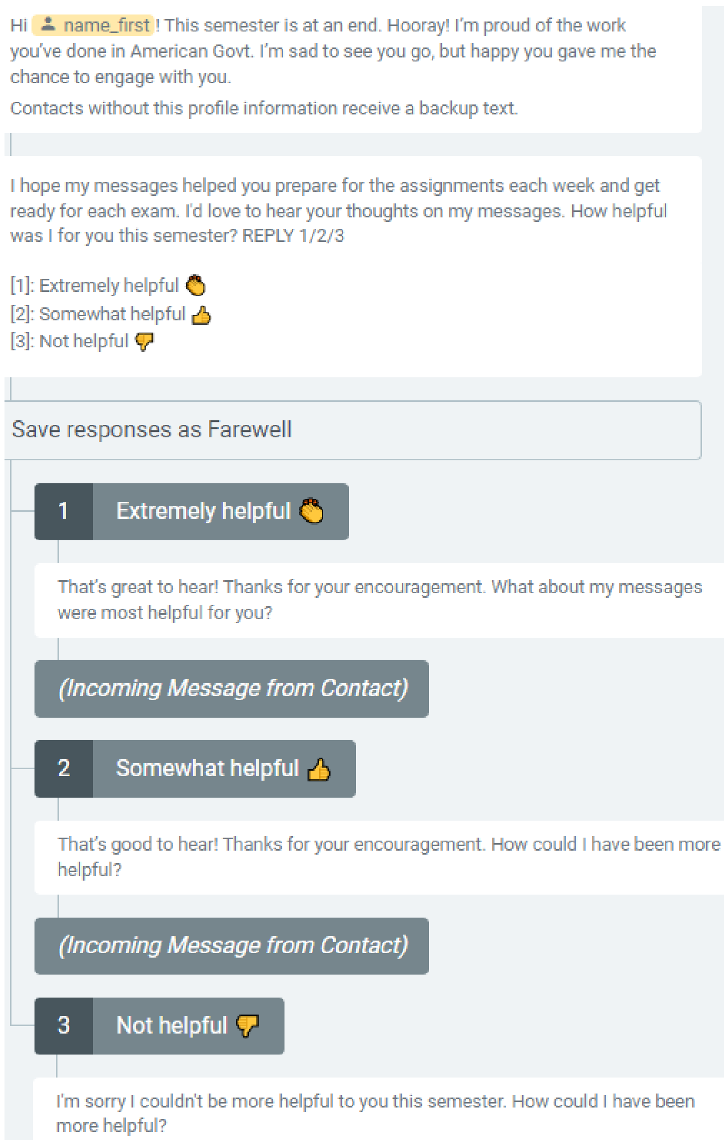
So that we can better support our students, would you mind sharing what went well and what didn't go so well during studying or on the exam? Is there anything Dr. Evans can do to help support you on exams?

3 Not so good. 😒

So that we can better support our students, would you mind sharing what didn't go so well during studying or on the exam? Is there anything Dr. Evans can do to help support you on exams?

FAREWELL INTERACTIVE MESSAGE_12.13.2021

Department / Office	Political Science 1101
Purpose	Farewell message wishing them well and asking for their quick feedback on how helpful the bot was for them this semester.
Target Population	225
Successful Contacts	207
Script	



APPENDIX C – Attitudinal Survey Measures

Organizational Support (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I know how the new things we're learning in this class connect to what we've learned before.
- This instructor regularly checks in to make sure we understand the class material.
- I feel like this class is organized to help me do well.
- It's clear what we're supposed to be doing in this class.
- I can communicate with this instructor about this class as needed.

Institutional Growth Mindset (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- This instructor seems to believe that students have a certain amount of intelligence, and they really can't do much to change it.

Self-Efficacy (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I have felt confident about my ability to do well in this class.

Inspiring Expectations (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I feel like this instructor trusts I can persist through challenging course material.
- I feel like this instructor thinks I can learn anything that is taught in classes.
- I feel like this instructor expects I will keep improving as a student.
- I feel like this instructor believes I have real potential in school.
- I feel like this instructor sees me as someone who could be successful in academics.
- I feel like this instructor recognizes that I can earn good grades if I put the effort in.

Adaptive Student Attributions (1 = "Not at all likely"; 5 = "Extremely likely")
If the following situation occurred during this course, how likely is it that you would have the thoughts below?

- You have to miss an exam for personal reasons.
    - I would think, "This instructor will be inflexible or unsupportive"
    - I would think, "This instructor will be understanding and helpful"
- You fall behind on the coursework one week, and the instructor messages you to say they noticed you still needed to turn things in.
    - I would think, "The instructor thinks I don't care about my education"
    - I would think, "The instructor is concerned about how I'm doing"
- You are doing poorly in the course and are at risk of failing.
    - I would think, "The instructor probably thinks I should drop the course."
    - I would think, "The instructor probably thinks I can pick my grade up."

GSU Challenge/Threat Ratio (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I feel like GSU will be a positive challenge for me.
- I feel like I have what I need to be successful at GSU.

- I am worried that some of the work at GSU will be stressful or overwhelming. (reverse-coded)
- I am uncertain if I could perform well in future GSU courses (reverse-coded)

GSU Social Belonging and Belonging Uncertainty (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I feel like I belong at GSU.
- I feel comfortable in classes at GSU.
- I feel accepted at GSU.
- I feel like I can be myself at GSU.
- Sometimes I feel that I belong at GSU, and sometimes I feel that I don't belong. (reverse-coded)

Trust and Fairness (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- This instructor treats me with respect.
- I trust this instructor to treat me fairly.
- I feel like the instructor truly has the best interest of their students in mind. (Eric added)

Meaningful Work (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- In this class, we do meaningful work, not busy work.
- What we learn in this class is connected to real-life.
- This teacher makes what we're learning really interesting.
- I feel like the course material to be relevant or useful to my life.
- I have been able to connect the course material to my interests or values.

Nervousness with Instructor (1= "Not at all nervous"; 5= "Extremely nervous")

- Imagine you decided to meet one-on-one with the instructor.
    - How nervous would you be about meeting this instructor?
    - How nervous would you be about having something to talk about?
    - How nervous would you be that they might judge you if you ask a "dumb" question?

Teacher Caring (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I feel like this instructor is glad that I am in their class.

Broad Regard (1 = "Strongly Disagree"; 6 = "Strongly Agree")

- I feel like this instructor would like to learn about my life outside of school.
- I feel like this instructor cares about what I do outside of my coursework.
- I feel like this instructor recognizes I have many identities beyond being a student.
- I feel like this instructor sees me as a person with many goals and values.
- I feel like this instructor welcomes my personal background and history.
- I feel like this instructor appreciates that I spend time on interests outside of schoolwork.