Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1083

Examining the Relationship Between Randomization Strategies and Contamination in Higher Education Interventions

Catherine Mata
Brown University

Katharine Meyer
Brookings Institution

Lindsay Page
Brown University

Randomized controlled trials (RCTs) are the reference method for causal inference. To conduct field experiments in educational settings, study design must balance statistical power with the risk of treatment-control contamination. This study investigates both crossover and spillover contamination in a large-enrollment, in-person college course in which we tested an AI-enabled chatbot intervention. We compare two randomization approaches, individual-level and laboratory-level, to assess contamination risks. Contrary to expectations, no crossover occurred under student-level randomization. However, survey data indicate evidence of spillover, with treatment-group students reporting that they shared chatbot messages with peers. Using estimated contamination levels, we assess changes in minimum detectable effect size (MDES) and show that individual-level randomization remains preferable. Our findings offer practical guidance for balancing contamination risk and statistical power when designing RCTs in interactive educational settings.

VERSION: July 2026

Suggested citation: Mata, Catherine, Katharine Meyer, and Lindsay Page. (2026). Examining the Relationship Between Randomization Strategies and Contamination in Higher Education Interventions. (EdWorkingPaper: 24-1083). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/rq74-c249

Examining the Relationship Between Randomization Strategies and Contamination in Higher Education Interventions

Catherine Mata
Brown University

Katharine Meyer
Brookings Institution

Lindsay Page*
Brown University

June 05, 2026

Abstract

Randomized controlled trials (RCTs) are the reference method for causal inference. To conduct field experiments in educational settings, study design must balance statistical power with the risk of treatment-control contamination. This study investigates both crossover and spillover contamination in a large-enrollment, in-person college course in which we tested an AI-enabled chatbot intervention. We compare two randomization approaches, individual-level and laboratory-level, to assess contamination risks. Contrary to expectations, no crossover occurred under student-level randomization. However, survey data indicate evidence of spillover, with treatment-group students reporting that they shared chatbot messages with peers. Using estimated contamination levels, we assess changes in minimum detectable effect size (MDES) and show that individual-level randomization remains preferable. Our findings offer practical guidance for balancing contamination risk and statistical power when designing RCTs in interactive educational settings.

JEL codes: C9, I23, D9

Keywords: contamination, randomization, causal inference, experimental design, higher education.

* Mata: catherine_mata@brown.edu; Meyer: kmeyer@brookings.edu; Page: lindsay_page@brown.edu. This research was supported through a grant from the Ascendium Education Group. We gratefully acknowledge Georgia State University (GSU), the National Institute for Student Success, and the Chemistry Department at GSU for their support and engagement with this study. We thank conference participants at the Society for Research on Educational Effectiveness 2024 conference, the Southern Economic Association 2024 annual meeting, and the Association for Education, Finance, and Policy 2025 conference helpful feedback. All errors are our own.

Examining the Relationship Between Randomization Strategies and Contamination in Higher Education Interventions

Randomized controlled trials (RCT) are the method of reference in quantitative causal inference (Angrist & Pischke, 2009, 2015; Murnane & Willet, 2011). In higher education, RCTs frequently are used to evaluate interventions related to financial support, enhanced advising, and tutoring (Diamond, 2025). Recent growth of artificial intelligence (AI) in education settings has led to an increase in randomized trials evaluating AI's impact on learning and information retention, as well as descriptive studies of students' use and acceptance of AI (Barcaui, 2025; Greenspan, 2025; Kestin et al., 2025; Owan et al., 2025). RCTs have also been employed in the higher education context to evaluate the effects of AI-enabled chatbots on mental health outcomes, academic performance, and college persistence (Fitzpatrick et al., 2017; Fulmer et al., 2018; Lambert et al., 2026; Nurshatayeva et al., 2021; Xu & Ma, 2025).

Such field experiments in educational and other social settings involve non-trivial design decisions. For instance, when implementing experiments in settings such as schools and classrooms where students constantly interact, a key decision is whether to randomize at the individual or cluster level. This decision involves weighing the tradeoff between statistical power and estimation precision, which are improved with individual-level randomization, and the risk of contamination, which is reduced with cluster-level randomization (Plewis & Hurry, 1998; Shadish et *al*., 2002; Bloom, 2005; Rhoads, 2011). Such contamination can attenuate the estimated treatment effect (Torgerson, 2001; Rhoads, 2011).[1]

[1] Another critical factor in whether to randomize at the individual or cluster level is the intervention itself and the level at which the intervention should be implemented. For example, interventions that treat the students could be randomized at the student or class level, whereas interventions that treat the instructor could only be randomized at the instructor or higher clustering level. For the sake of our exercise in this paper, we only consider the case where the intervention could be randomized at the individual or cluster-level, and the researcher must weigh the tradeoffs associated with randomization at these levels.

Previous work has discussed the risk of contamination when implementing individual-level randomization in educational and other social settings (Plewis & Hurry, 1998; Shadish et *al.*, 2002; Bloom, 2005; Rhoads, 2011). This contamination can take two different forms. First, control-group students can actively seek and gain access to the treatment. This kind of contamination is called crossover. Second, control-group students' outcomes can be influenced through exposure to their treated peers. This is called spillover, given the notion that the effects of an intervention could spill over from the treated students to those in the control condition. Whereas crossover can be observable, depending on the extent to which the researcher can control or observe access to the intervention, spillover can be harder to observe or measure directly.

At first glance, it may seem logical for a study designer to aim to avoid contamination at all costs, given its potential to attenuate treatment effect estimates. However, Rhoads (2011) demonstrates that, given the precision loss associated with cluster-level random assignment, individual-level randomization with some contamination can still be more powerful and, therefore, preferable to cluster-level randomization with none. A key question, then, is how much contamination can be expected and tolerated in a given context. In this paper, we report on our process for considering this question in the context of designing a specific experimental study.

We first measure the extent of control-group crossover contamination across two different randomization schemes that we piloted to assess the feasibility of estimating the impacts of our focal intervention in a large-enrollment, in-person college course. We then conduct an exploratory analysis of spillover contamination using end-of-course student surveys. We use the findings from these surveys to estimate the likely extent of treatment-to-control spillover under simple individual-level randomization—the preferred randomization scheme identified in our first step. Finally, we assess the implications for statistical power based on the extent of crossover and

spillover contamination that we estimate. We walk through these steps in service of two goals. First, we document our understanding of contamination in our experimental setting. Although aspects of this context may not generalize, the challenges and decisions we encountered are likely to have broader applicability. Second, we aim for this paper to serve as a guide for others weighing similar tradeoffs between individual- and cluster-level randomization in other contexts and with other interventions. In this sense, the specific tools and strategies that we use for gauging crossover and spillover may not generalize to other contexts, but the thought process and steps that we walk through likely do.

Building on Rhoads' (2011) foundational work, we examine the levels of contamination in an experimental intervention set in a higher education context. Our experimental intervention involves course-specific, text-based, AI-enabled chatbot communication to provide students with regular outreach, encouragement, and reminders about available academic supports on their college campus. We present empirical evidence on the level of control-group crossover encountered in different approaches to randomizing students to the intervention implemented in a large-enrollment, in-person, undergraduate course. Across two lecture sections with nested, required, weekly laboratory sessions, we piloted two different approaches to randomization: (1) student-level randomization of students attending the same large lecture and (2) within-lecture section, cluster-level randomization, in which all students within a laboratory subsection were randomized either to treatment or control. Prior to the start of the semester, we established that any control group student who learned about and requested chatbot communication would be afforded this access. Nevertheless, by the end of the first semester of implementation, no control-assigned students requested access to the chatbot communication under either randomization scheme.

Given its large lecture setting, communication and socializing may be challenging for students in this course. However, students also spend three hours each week in their smaller laboratory groups. We were particularly surprised that, even with student-level randomization that led to treatment and control students being in the same laboratory sections together, no control-assigned students crossed over into the treatment condition. To ensure the robustness of our findings, we implemented student-level randomization in three additional lecture sections during the subsequent semester. We again observed no instances of students in the control group entering the treatment condition in any section.

Nevertheless, data from an end-of-course survey suggest some spillover contamination. Specifically, in their survey responses, some treatment-group students reported sharing chatbot messages with their classmates. While only 10% of control-group respondents indicated receiving chatbot messages or related information from peers, approximately 30% of treatment-group respondents reported sharing such content. Whether control-group students recognized the source of information shared and whether recipients of shared information from treatment-group members were in the control group are unclear. Nonetheless, these findings suggest the potential for spillover in student-level randomization as this intervention continues.

To assess how contamination affects the minimum detectable effect size (MDES) in our study, we use the observed spillover levels in our setting (10 to 30 percent) to estimate the resulting changes in MDES. As we detail below, at these spillover levels, student-level randomization remains preferable to laboratory-level randomization. Furthermore, assuming a scenario in which spillover affects control units as intensely as direct crossover, we find, consistent with Rhoads

(2011), that contamination would need to affect more than half of the control group before laboratory cluster-level randomization would be the preferred design.[2]

These results give us confidence to proceed with our own experiment with an individual-level randomization study design. This approach and decision-making process offer a useful reference point for others implementing experimental interventions in comparable settings. More broadly, our application provides an example for researchers and practitioners of how to investigate potential spillover between treatment and control groups using survey data and how to use estimated spillover rates to update MDES in subsequent impact analyses.

Intervention, context, and level of randomization

The intervention itself and the context in which it is implemented affect the freedom the researcher has for choosing an appropriate level of randomization. The first determination a researcher must make is what unit (e.g., student, teacher, school) the intervention is targeting and what possible levels of randomization align with their theory of change. While some interventions that operate at the individual level can be randomized either at the individual or cluster level, interventions that directly affect the cluster altogether, of course, cannot be randomized at the individual level. For example, treating students with a new program may leave some room to structure randomization such that some students are randomized to the treatment conditions and some classmates to the control condition while assessing for crossover and spillover. However, if

[2] We recognize that the risk from contamination arises not only from its magnitude but also from its composition. Selection into contamination can bias treatment effect estimates. For example, control-group students who are exposed to the treatment may be systematically more or less likely to benefit than the average student. In this case, contamination does not simply attenuate the treatment-control difference; it can also introduce bias. If researchers observe crossover, they can use instrumental variables to estimate the local average treatment effect (LATE) of treatment on outcomes, defined with respect to the assignment mechanism. However, this parameter might not coincide with the average treatment effect (ATE), particularly in the presence of non-random always-takers. While these issues are important, they are beyond the scope of this paper. We instead focus on the design stage, where researchers must choose the level of randomization to maximize statistical power, often in settings where detailed data on crossover are not available for two-stage estimation. In such cases, researchers may only have access to an approximate range of crossover, as in our survey-based evidence.

the focal treatment targets their teacher, the class section (cluster) becomes the unit of randomization, and all students (individuals) within the cluster are assigned to the same treatment status as their teacher.

The intervention we study is a communication tool for students, and our setting is a large, in-person college course. This setting allows us to choose between individual-level and section-level (cluster) randomization. Although individual-level randomization offers greater statistical power for impact estimation, we judged potential contamination as among the largest risks to the success of our study, and therefore deemed it necessary to assess the risks of contamination arising from crossover and spillovers in a pilot context before launching the full study in multiple sections of our focal course.

Our work emerges from a research-practice partnership with Georgia State University (GSU) to understand the effectiveness of incorporating AI-enabled chatbot communication into systems of outreach and support for the university's undergraduate population. GSU uses a chatbot built by the technology company Mainstay and has been testing and expanding the use of the tool in different aspects of student communication and support over time.[3] GSU is a large, public university located in Atlanta, GA, that serves a diverse student population.

Previous work (e.g., Page & Gehlbach, 2017; Meyer et *al*., 2024; Page et *al*., 2025) has reported on the positive effects of text-based chatbots with AI capability in supporting students to navigate administrative processes and use campus resources, as well as on students' academic task navigation and performance in introductory political science and economics courses. In these course-specific studies, both courses were offered online and asynchronously, so the students enrolled in them seldom, if ever, interacted with each other in person or online. GSU is now testing

[3] For more information on Mainstay, please see https://mainstay.com/.

the tool's effectiveness in Chemistry 1211 (CHEM1211), the university's first-semester chemistry course for STEM majors. Different from the prior courses in which the chatbot was tested, CHEM1211 is taught in person. Given this in-person context, we reasoned that we needed to contend with an increased potential for control-group crossover and spillover and so piloted different randomization structures to investigate the associated prevalence of contamination within them.[4]

In the context of an in-person class, the contamination that might come with individual-level randomization creates multiple concerns. Both crossover and spillover lead to control group exposure to the intervention. In this way, both could lead to an attenuation of the estimated treatment effect through upward bias of the control-group mean (assuming a positive treatment effect). Further, especially given the in-person setting, we hypothesized a scenario in which faculty members would be put in the situation of having to answer to control-group students who learned about the chatbot and were not initially given access. Before expanding the experimental intervention to more course sections, we aimed to understand how prevalent crossover and spillover were likely to be.

Here, we report on the two randomization schemes we piloted for studying the chemistry course chatbot in CHEM1211. In the Spring 2024 semester, students in two sections of CHEM1211 who, upon enrollment in the university, had agreed to receive university communication via text message (i.e., "text-eligible" students) were randomized either to control

[4] The impact evaluation of this intervention is pre-registered with the Registry of Efficacy and Effectiveness Studies (REES) under Registry ID 20641, and the exploration of the best approach to randomization is pre-registered under Registry ID 18140. The Institutional Review Boards (IRBs) at Georgia State University and Brown University reviewed and approved the impact evaluation of the chatbot intervention, including the administration of the end-of-course survey used for this study and the broader research project.

or treatment (see Figure 1).[5] Across both lecture sections, students participated in large lectures of approximately 200 students each and small laboratory subsections of approximately 25 students each. Two instructors, referred to here as Instructor A and Instructor B, each with over a decade of experience teaching this course, agreed to collaborate on the intervention study. Students enrolled with Instructor A were randomized at the student level.[6] Approximately half of the students in Instructor A's lecture section were assigned to receive text communication from the chatbot, and the other half were assigned to a control condition receiving regular course communication but no text-based outreach.[7]

Text-eligible students enrolled with Instructor B were randomized at the laboratory level, with randomization nested within laboratory instructor. At the time of this pilot, students enrolled in the same course section were subdivided into eight laboratory sections, taught by four different laboratory instructors, each of whom oversaw two laboratory sections.[8] Within each of the four laboratory instructors, one laboratory section was randomly assigned to treatment, and the other was randomly assigned to control.

Few students added or dropped the course after our initial randomization; however, we repeated the randomization process with students who enrolled in the class after it began but before the end of the add/drop period. The total analytic sample includes 192 students assigned to

[5] Upon enrollment at GSU, students may opt in to receive text-based communications from the university, including the retention chatbot Pounce. In our sample, fewer than 7% of students opted out and were therefore classified as "text-ineligible" for the academic chatbot studied here. Text-eligible students are broadly comparable to their ineligible counterparts, with minor differences in certain semesters: Pell-eligible, first-generation, and first-year students are somewhat more likely to opt in. These differences, however, are modest.

[6] We implemented individual-level randomization at the lecture level, using blocks defined by laboratory sections to maintain a near 50/50 split between treatment and control within each laboratory section.

[7] Regular course communication includes announcements in the course platform, in-class verbal communication, emails, and office hours.

[8] Since we piloted in CHEM1211, GSU changed this laboratory structure. In the new structure, laboratories are not nested within faculty lectures anymore, and instead, students enroll in a lecture and laboratory separately. Therefore, we cannot test this same randomization scheme in subsequent semesters.

treatment and 170 students assigned to control, nearly evenly split between student-level and laboratory-level randomization approaches. Please see Figure 1 for complete details of the randomization.

Students assigned to the treatment condition received regular outreach and support with specific course content and general academic competencies via text messages. On average, students received three messages per week over the course of the 17-week-long semester: a weekly digest or routine communication each Monday, a targeted or interactive support message midweek, and an interactive quiz tool (#*quizme*) by the end of the week. Chatbot content addressed three primary domains within which previous cohorts of students have reported difficulty: (1) time management; (2) academic course content; and (3) chemistry belonging. See Figure 2 for examples of the text message outreach (called "campaigns") students received throughout the semester. Students could respond via text message to ask questions at any time. These questions were answered either immediately by the chatbot AI or as soon as possible by a course teaching assistant in cases where the AI could not adequately answer the question. If students decided they did not want the text-based outreach, they could pause or opt out of the chatbot communication at any time by messaging PAUSE or STOP to the chatbot.

The chatbot was mentioned neither on the course syllabus nor by faculty during class sessions or office hours. Nevertheless, before the semester began, we planned that any control group student who learned about the chatbot from peers and requested access could also receive full access to it, including the proactive campaigns. Faculty were instructed to inform interested students that this was a new GSU program and that they could participate by requesting access via email. While faculty did not actively encourage students to request access, systems were in place

to ensure that any student who expressed interest and requested access would be added easily and promptly.

The work on which we report here is part of a larger, multi-semester effort to assess the impact of the chatbot on students' academic outcomes, including final course grades, engagement with academic supports such as tutoring, performance on coursework (homework, quizzes, and exams), and progression to the subsequent course in the chemistry sequence. These impact analyses are ongoing and beyond the scope of our purpose and reporting here.

Data and Methods

Data

We rely on three sources of student-level data to assess chatbot system engagement among treated students and intervention contamination among control students: (1) administrative records held by the university, (2) chatbot engagement metrics from the messaging platform, and (3) end-of-course survey responses.

Student-level administrative data: We use student-level characteristics captured in university administrative data, including race/ethnicity, sex, first-generation status, level of financial aid received, enrollment intensity (e.g., full- or part-time), year in college, and high school or university GPA. These variables serve as baseline measures at the time of randomization. We use these measures to assess balance between students assigned to the treatment and control groups and to consider post-randomization variation in behaviors and actions, such as chatbot engagement, opt-out, and control group crossover.

Student-level engagement with the chatbot: Using de-identified records from the Mainstay platform, we observe each outreach message sent from the chatbot to students and each response from students to the chatbot. These data enable us to assess several aspects of chatbot usage. First,

we can observe if and when students opt out of receiving messages from the course chatbot. Likewise, we can observe if and when control students opt into chatbot communication, allowing us to track the crossover of control students into treatment. Second, we can gauge active student engagement by tracking the number of messages students send to the chatbot.

End-of-course survey: In Fall 2024, we administered an end-of-course student survey in one of the CHEM1211 sections, where the instructor incentivized participation by offering extra credit for completion.[9] This survey implementation design yielded a 60 percent response rate and a well-balanced sample of respondents who are not systematically different from those who did not complete the survey. As shown in Table 1, respondents and non-respondents are similar across a wide range of covariates, including current and prior exposure to the intervention, race/ethnicity, age, sex, socioeconomic status, first-generation status, and various measures of academic performance. The only notable difference is in college GPA, with students who responded to the survey having, on average, a GPA 0.55 points higher than that of non-respondents, despite both groups having similar high school GPAs.

The survey was tailored to the students' experimental group assignment. From treatment-group students, the survey was designed to gather insights into whether and how chatbot communication was a valuable aspect of their course experience and whether they shared chatbot messages or information coming from the chatbot with any classmates. Students were reassured that it was completely fine if they shared chatbot messages or information with classmates and that we were just trying to understand how the chatbot was used in their course. From control-group students, the survey was designed to gather insights into the reasons why they might have been more or less likely to request access to the chatbot. Students in the control group were asked

[9] We attempted to administer the same end-of-course survey in a previous semester without offering extra credit for completion, but the student response rate was too low to yield data suitable for analysis.

whether they had heard about the chatbot from their classmates. These data inform whether control students (1) were aware of the existence of the chatbot in their class, (2) received chatbot information through their peers, in turn, reducing the need to opt into the chatbot, or (3) were not interested in the chatbot despite knowing of its existence. Responses to these questions inform whether there is evidence of spillover from students in the treated group to students in the control group.

Methods

In our analysis, we first focus on the Spring 2024 data to investigate the extent to which control-group students sought formal access to the chatbot and to compare the levels of crossover between the two randomization schemes. We analyze administrative data to explore potential explanations for the patterns of control-group crossover. We then test the robustness of the first semester results using a second semester of data with individual-level randomization in three lecture sections. In the Fall 2024 semester, we continued testing the chatbot communication and implemented student-level randomization across three lecture sections of CHEM1211 taught by the same instructors. Instructor A taught two sections with a combined total of 380 students, 361 of whom were text eligible. Instructor B taught one section with 182 text-eligible students. Across all three lecture sections, 280 students were assigned to treatment and 263 to control. Please see Figure 3 for further details on the Fall 2024 randomization.

Using the end-of-course survey data from the second semester, we investigate evidence of spillover and consider potential explanations for the lack of formal crossover we observe. Finally, we use our estimates of crossover and spillover to calculate the potential statistical power loss associated with the levels of contamination found and consider whether the contamination due to crossover or spillover affects the design preference for individual-level randomization.

Results

Crossover contamination

During Spring 2024, students assigned to the treatment group received text-based communication throughout the entire academic semester. Surprisingly, during this time, none of the control-group students—whether in the individual- or laboratory-level randomization scheme—requested access to the course chatbot. We again found no crossover in the Fall 2024 lecture sections, within which we used student-level randomization. These clear and consistent results have alleviated concerns about student-level randomization in large lecture courses contributing to treatment effect attenuation or placing faculty in potentially awkward situations with regard to control-group students asking for chatbot access in large numbers.

We considered several possible explanations for the lack of control group crossover. First, students may be generally uninterested in receiving this type of course-specific outreach. To gauge receptivity and interest, we consider passive engagement (via opt-out rates) and active engagement (via the share of students who messaged into the system and the number of messages that they sent) among students in the treatment group. Opt-out was rare; only 11 treatment-group students (5.7 percent) opted out in Spring 2024, and 8 treatment-group students (1.5 percent) opted out in Fall 2024. In the Fall 2024 semester, 42.5 percent of treatment-group students sent at least one message to the bot, and students who replied to the chatbot at all sent an average of 6.5 messages. Active engagement was driven, in part, by the practice quiz function – about 44 percent of students

who ever replied engaged with the quiz function.[10] In summary, treatment students were highly engaged with the chatbot, as evidenced by both low opt-out rates and high active participation.[11]

To further investigate perceptions of the benefit and value of the course-specific chatbot communication, we analyzed students' responses to the end-of-course survey conducted in Fall 2024. In prior studies of GSU's course chatbots, students generally reported enthusiasm for the tool, with over 90 percent of survey respondents recommending its continued use in the course and expansion to other courses at GSU (Meyer et *al.*, 2024). As noted above, these prior studies were situated within courses taught online rather than in person. In the in-person course context, students may feel less positive or may perceive less of a need for chatbot communication. Of the Fall 2024 treatment students who responded to the survey, over 90 percent reported reading the course chatbot messages, and 74 percent found the messages helpful or very helpful. Most of the remaining students were neutral about the message's helpfulness, with only 5 percent of respondents reporting not finding any value in the texts. Over 93 percent of treatment-group respondents recommended continuing to use the chatbot in the same chemistry course, and 88 percent recommended expanding its use to other courses.

Second, we hypothesized that lack of crossover may be driven by weak social connections among peers and the lack of an opportunity for control students to learn about the chatbot because they did not experience sustained conversations with classmates. Despite the value treatment-group students found in the chatbot communication and greater potential for crossover in an in-

[10]Active engagement rates were higher in the spring 2024 pilot, when the bot more actively promoted the quiz function. During spring 2024, 69 percent of treatment-group students sent at least one message to the bot, with 59 percent of students who ever replied engaging with the quiz function.

[11] As prior studies show, engagement with these types of tools can be measured along two dimensions: the share of students who actively respond to the chatbot (i.e., active engagement) and the share who remain subscribed and receive messages without interacting (i.e., passive engagement). The latter is typically proxied by lower opt-out rates. Active engagement rates vary substantially depending on the intended communication of the tool, with reported ranges of 5-85%, as documented by Page and Gehlbach, 2017 (85%) and Mata et al., 2026 (5%) for admission and retention chatbots, and Meyer et al., 2024 (49%) for other academic chatbots.

person course context, among control-group survey respondents, only 10 percent (a total of 5 students) reported knowing about the existence of the course chatbot. This is consistent with treatment students not explicitly sharing the chatbot communication with their peers. This lack of sharing could be consistent with weak social connections among students in the course or with low appreciation for the intervention among treated students, the latter of which we have already ruled out. Based on our survey data, control students reported having connections with class peers. Seventy-seven percent of all control-group survey respondents indicated interacting with between one and six or more classmates outside of class meetings.

In sum, formal control group crossover was surprisingly non-existent in the course context we investigated. Based on findings from system usage data and the end-of-course survey, we conclude that treatment students found value in the course-based chatbot communication and that the lack of control-group crossover was not due to a lack of treatment-group student interest or receptivity or to a lack of social connections among the students in the course. Yet another possibility is that the lack of crossover is driven by a high degree of spillover. That is, if control-group students were exposed to the content of chatbot communication from their treated peers, then they may have considered formal access to be unnecessary, or, more likely, they may not have recognized the source of the information that their treated peers were passing along. We explore evidence of such spillover in the next section.

Spillover contamination

Of the 10 percent of control group respondents who reported knowing classmates who received text messages from the course chatbot in our focal course, only one student indicated that a classmate had shared chatbot messages or information with them.[12] However, among treatment

[12] Given the limited number of control-group students who reported awareness of the intervention, we do not report their characteristics, as no meaningful conclusions about spillover patterns can be drawn.

group respondents, 30 percent (a total of 17 students) reported sharing chatbot messages or the information that the messages contained with their classmates. Of course, if this sharing occurred in a balanced way with treatment- and control-group students, it would imply a spillover rate of about 15 percent. These results have several implications for our specific context and beyond. First, within this study's context, spillover is likely to have occurred through treatment students sharing information from the chatbot messages without necessarily mentioning the mode of communication itself, and control students may have gained information originating from the chatbot without being aware of its source. More generally, our pattern of findings underscores that spillover contamination may be occurring, even in the absence of formal crossover. Such spillover without crossover may occur among control-group students because of a lack of interest in the intervention, a lack of awareness of the intervention, or a lack of awareness of how to access the intervention. Even if aware, control-group students may not have considered it possible to ask their course faculty for access to the tool.

Taken together, our findings underscore the importance of assessing spillover contamination, even when there is no direct evidence of crossover contamination. As our case also illustrates, a survey implemented during the pilot or early stages can help identify spillover when other mechanisms for observing spillover are unavailable. While such surveys may not fully capture information sharing, they can still inform estimates of spillover magnitude to inform subsequent design decisions. In the next section, we turn to the technical details of how the spillover we detect affects the statistical power of our experimental study.

Individual randomization with contamination

A logical next step for our analysis is to determine whether individual-level randomization remains preferable to cluster-level randomization in the absence of crossover but in the presence

of spillover contamination. To this end, we focus on calculating the Minimum Detectable Effect Size (MDES) under different scenarios. That is, instead of asking, "What is the power to detect a given effect size?" we focus on the question, "What is the smallest effect we could detect with 80 percent power, given our sample?" In general, calculating the MDES may be more practical because researchers often have fixed sample sizes due to budget, recruitment, or logistical limitations.

We refer readers interested in a formal derivation of MDES to Glennerster & Takavarasha (2013). Examining these formal derivations is useful for understanding how power, MDES, variance, and sample size are related and can inform experiment designers in making decisions that lead to more powerful studies. For calculations of the MDES, evaluators also can use preprogrammed packages in statistical softwares such as Stata (e.g., *sampsi*, *sampclus*, *power*) and R (e.g., PowerUpR) or other available tools such as *PowerUp* developed by Dong et al. (2015), which easily adapt to various experimental designs and allow users to input their specific parameters to obtain the MDES without needing to program the formula.[13]

Given its ease of use, we will use *PowerUp* to walk through the steps of our MDES calculation. When using any preprogrammed tool, however, it is important to understand its underlying assumptions. For instance, in *PowerUp*, when inputting an assumed R^2, researchers must ensure it is derived from a standardized outcome so $\text{Var}(Y) = 1$ and $\sigma^2 = 1 - R^2$.

PowerUp provides a table of contents with different study designs and their corresponding MDES calculation spreadsheets. In our case, we used model 2.1 for blocked individual random assignment, and model 3.1 for simple cluster assignments with two levels of clustering. In the

[13] References for software implementations are as follows: Stata packages (Glennerster & Takavarasha, 2013, Ch.6), R packages (Ataneka, 2023; Bulus, 2022), and the PowerUp tool (Dong et al., 2015).

context of our initial experiment, students enrolled in a chemistry lecture course were subdivided into smaller sections for laboratory sessions, allowing for two different potential levels of randomization – either randomizing students (nested within laboratories) or randomizing laboratory clusters (nested within lecture sections).[14] We estimate the MDES under these two different randomization designs with eight lectures offered: (1) cluster RCT with two levels with randomization at the laboratory level, and (2) individual-level randomization (which could be completely random within a lecture or, as we opted for, blocked randomization within laboratory sections). In the first column of Table 2, we present the MDES for the cluster randomization at the laboratory section level. In this case, we assume that half of the clusters are assigned to treatment, that individual covariates explain about 20 percent of the variance in the outcome and cluster covariates another 2 percent (Bloom et al., 2007), that intraclass correlation is approximately 10 percent (Hedges & Hedberg, 2014, Bloom et al., 2007), and there is no contamination.

Randomizing at the laboratory section level, as we implemented with one of the course sections during our Spring 2024 pilot, renders an MDES of 0.25. This means that with a cluster-level randomized design with clusters defined at the laboratory level, we should be able to detect effects that are 0.25 standard deviations or greater. This MDES is considered a large effect for most education interventions studied in field settings (Kraft, 2020).[15]

In column 2 of Table 2 we present the MDES estimate for individual-level randomization (blocked at the laboratory level) under no contamination. Unsurprisingly, in the absence of contamination, individual-level randomization is preferable to cluster-level randomization. This is consistent with theoretical expectations, given the superior statistical power associated with

[14] A third option was lecture-section randomization, which we deemed impractical given the small number of sections and the resulting lack of statistical power to detect meaningful effects.

individual-level randomization under ideal conditions. Without contamination, individual randomization yields an MDES of 0.13—half the size of the MDES under laboratory-level randomization. This means that, given our sample, randomizing at the individual level would allow us to detect effects of 0.13 standard deviations or greater. Moreover, an MDES of 0.13 falls near the midpoint of what Kraft (2020) identifies as medium-sized effects in experimental education research. For the intervention at hand, we expected small to medium effect sizes based on prior evaluation of course chatbots in different settings. For instance, Meyer et al. (2024) found that students assigned to a chatbot in online classes were about four percentage points more likely to earn an A or B in the course, representing an effect size of 0.09. While we were interested in maximizing our capacity to detect small to medium effects, in other intervention contexts where the researcher hypothesizes the effect of the treatment to be larger, the reduction in MDES from using cluster-level randomization may be of less consequence.

We now relax our previous assumption of no contamination to allow for the possibility that control-group students may be exposed to the intervention through their peers in the treatment group. In our MDES calculations, we take a conservative approach by treating spillover and crossover as comparable forms of contamination. In reality, spillover likely involves partial rather than full exposure to the intervention. Following Glennerster and Takavarsha's (2013) practical guide for randomized evaluations we adjust the MDES for partial compliance as follows,

$$MDE_{\substack{partial \\ compliance}} = \frac{MDE_{full\ compliance}}{P_t - P_c}, \quad (1)$$

where P_t is the proportion of the treatment group that is treated and P_c is the proportion of the control group that is treated (contamination rate). Because the MDES calculation in previous steps

already accounts for the randomization structure, the formula in (1) does not require further adjustment. This formula can be applied to adjust for partial compliance after estimating the MDES under full compliance, regardless of the RCT structure.

We use the results from the previous section to assess how the MDES would change if we were to continue with individual-level randomization under spillover levels similar to those observed during our pilot semesters. Specifically, we consider two scenarios in which individual-level randomization is used and spillover contamination affects 10 and 30 percent of the control group, respectively. These percentages reflect a plausible range for the rate of spillover based on our survey data.

Following Rhoads (2011), we treat all forms of contamination equivalently. That is, we assume that spillover contamination has the same impact on the estimated treatment effect as crossover. In the absence of actual crossover (as confirmed in our pilot data), we treat spillover cases as if they were fully contaminated, equivalent to participants who directly received the treatment. In reality, spillover likely results in partial rather than complete treatment exposure, making 30 percent a conservative upper bound for our estimated rate of contamination.

When accounting for the possibility of a 10 percent contamination rate in the experiment under individual randomization, the MDES with partial compliance remains close to that of full compliance or no contamination, with an MDES of 0.14.[16] Using the upper bound of 30 percent contamination, we estimate an MDES of partial compliance of 0.19. This is still more favorable than randomizing at the laboratory level.

This same exercise can be repeated for individual-level randomization under different assumed levels of contamination, generating the corresponding MDES for each scenario. These

[16] $MDE_{c=10\%} = \frac{0.13}{1-0.1} = 0.14$

estimates can then be plotted along the MDES from cluster-level randomization without contamination, given the sample size of the experiment. We illustrate this exercise for our setting in Figure 4.

This comparison provides a straightforward way to identify the point at which cluster-level randomization becomes preferable to individual-level randomization. In our case, this occurs when the contamination rate exceeds 48 percent. To the left of this threshold in Figure 4, the MDES under cluster-level randomization is larger than the MDES under individual-level randomization, indicating that individual-level randomization provides greater statistical power and allows us to detect smaller effects. By contrast, once contamination exceeds 48 percent, the MDES under individual-level randomization becomes larger than that under cluster-level randomization, as reflected by the individual-randomization curve crossing above the cluster-randomization line. Beyond this point, cluster-level randomization would allow us to detect smaller effects and would therefore become the preferred design.

More generally, to formally find the contamination rate at which individual-level randomization with partial compliance (pc) would become less preferable than cluster-level randomization with full compliance (fc) we solve the following inequality:

$$MDES(cluster)_{fc} < MDES(indiv)_{pc}$$

$$MDES(cluster)_{fc} < \frac{MDES(indiv)_{fc}}{P_t - P_c}$$

$$P_t - P_c < \frac{MDES(indiv)_{fc}}{MDES(cluster)_{fc}}$$

$$P_t - \frac{MDES(indiv)_{fc}}{MDES(cluster)_{fc}} < P_c$$

In our case, we substitute $P_t = 1$ because all treated students received the intervention, and $MDES(cluster)_{fc}$= 0.25 and $MDES(indiv)_{fc} = 0.13$ from our calculations shown in Table 2, columns 1 and 2, respectively. We have that,

$$1 - \frac{0.25}{0.13} < P_c$$

$$0.48 < P_c$$

This means that, in our case, the control group contamination rate must exceed 48 percent for cluster randomization to yield a lower MDES than individual randomization with partial compliance. In other words, for individual-level randomization to become less preferable than cluster-level randomization, contamination would need to reach approximately 50 percent—that is, about half of the students originally assigned to the control group would have to gain access to the treatment either directly or through their peers.

These results are consistent with Rhoads' (2011) argument that when there are not too many clusters in the experiment (e.g., fewer than 100) and when the standardized effect size is not too large, then individual-level randomization will generally result in more precise estimates of the average impact of treatment and more powerful statistical tests of the null hypothesis of no treatment effect than the cluster-level design, even with a fairly large degree of contamination. In fact, Rhoads (2011) estimates that contamination can deteriorate the observable average treatment effect by between 10 percent and 60 percent before the cluster design becomes preferable to the individual randomization design.

It is clear, then, that in the context of our experimental study, individual-level randomization remains preferable to cluster-level randomization, even in the presence of some contamination. Throughout this analysis, we have assumed a fixed sample size, which is common in experimental settings due to budget constraints and other logistical limitations. However, when

there is evidence of potential contamination, such as spillover effects in our case, and there is room to increase the sample size, it becomes useful to ask: what sample size would be required to recover the MDES we would achieve under ideal conditions (i.e., individual-level randomization without contamination)? To respond to this question, Glennerster and Takavarsha (2013) recommend using the following equation,

$$N_{\substack{partial \\ compliance}} = \frac{N_{full\ compliance}}{P_t - P_c}, \quad (2)$$

For example, to calculate the sample size needed for our pre-established MDES of 0.13, we use the total number of students (N=1,600) as the sample under full compliance, P_t=1 because all treated students receive the intervention, and we set P_c=0.3 to incorporate the upper bound of 30 percent contamination rate. This results in a sample with partial compliance of about 3,265 students. This represents a little more than double the original sample size in the no-contamination scenario. This means that in our case, to return to an MDES of 0.13, we would need to increase the total number of laboratory sections to about 131 for a total of approximately 3,300 students. We model the possibility of increasing the number of laboratory sections within which we randomize students at the individual level to increase our sample size because, from a practical standpoint, this strategy is relatively more feasible, whereas increasing the number of students per laboratory section is not, given the physical constraints of the laboratory setting. However, even when there is some flexibility to increase the number of observations within clusters, it is important to remember that we could eventually exhaust the pool of students enrolled in this course and that adding more people per group at the expense of reducing the number of groups reduces statistical power. This happens because individuals within the same group tend to be more similar to each other, providing less variability in information (Glennerster & Takavarasha, 2013)

The broader takeaway for other researchers is that these steps can help determine whether individual- or cluster-level randomization is more appropriate in a given context, depending on expected contamination levels. When contamination is not readily observable in administrative data, self-reported survey data can yield useful estimates. Once the MDES under ideal conditions is known, researchers can use the estimated rates of contamination to reverse-engineer the necessary sample size to achieve that same MDES, even in the presence of contamination.

Conclusions

In this article, we build on foundational evidence showing that individual-level randomization is (1) a more efficient design for experimentally assessing program effectiveness, but (2) carries a higher risk of treatment-control contamination and resulting bias of the treatment effect estimate due to the potential crossover of control units into the treatment group and spillover of treatment effects from treated to control students who interact in social contexts (Bloom, 2005; Shadish et *al*., 2002; and Plewis & Hurry, 1998). The study on which we report here provides empirical evidence of the level of crossover encountered in two different approaches to randomization for an intervention implemented in the context of a large, in-person lecture course in college.

We pilot-tested a course-specific chatbot within two sections of a college-level introductory chemistry course. Within large lectures of almost 200 course enrollees, students were further subdivided into laboratory sections of a maximum of 25 students. In this context, we tested two different approaches to randomization: (1) student-level randomization and (2) laboratory-level randomization. To address ethical considerations, any control group student who learned about the chatbot and expressed interest in receiving the communication would have been allowed to opt into receiving the communication.

We hypothesized that the greatest risk for crossover would have occurred under the individual randomization approach, as this would lead to the scenario where both treatment- and control-assigned students were in small-group, in-person laboratory subsections together. We considered the laboratory-level randomization to be a strategy to mitigate this risk. To our surprise, we observed no crossover in either student- or laboratory-level randomization. Furthermore, an additional semester of data, with two more sections randomized at the student level, provides further support for this key result. Student chatbot engagement and survey responses suggest that the lack of crossover was not due to a lack of socializing among students in the course or a low appreciation of the chatbot tool among treated students.

However, survey responses suggest spillover effects, with 10 to 30 percent of treated students sharing information with others. Our estimates indicate that, given our sample, spillover would need to reach an intensity comparable to full crossover, where control units effectively receive the full treatment, and affect more than 50 percent of control-group students before laboratory-level randomization becomes preferable to individual-level randomization. This aligns with Rhoads (2011), who estimates that, depending on the sample size, contamination could affect between 10 and 60 percent of the sample before individual-level randomization becomes less desirable than cluster-level randomization.

The pilot study discussed here is part of a larger intervention study. In the first semesters, our primary focus was identifying a randomization approach that would mitigate the risk of contamination. While we recognize that student-level randomization with some contamination can still offer greater statistical power than cluster-level randomization (Rhoads, 2011), our goal was to estimate the potential level of contamination when randomizing at the individual level in an in-person class. The absence of any crossover and contained levels of spillover in the individual

randomization condition gives us confidence that the larger study can proceed with student-level randomization.

While the specific sample size, randomization design, and contamination rates in our setting may not generalize to all contexts, the thought process, tools and/or strategies we used to assess crossover and spillover, and the steps we followed in here can serve as a practical example for researchers facing similar design decisions. By walking through our approach, we aim to both inform understanding of contamination in our experimental setting and offer a resource for others weighing the trade-offs between individual- and cluster-level randomization. We also highlight that, in some contexts, even when individual-level randomization is statistically advantageous, ethical concerns and the risk of spillover, particularly in small or highly interactive learning environments, may warrant alternative designs. In such cases, withholding treatment from students aware of their peers' participation may not only compromise the study but also raise ethical and practical challenges for instructors.

References

Angrist, J. D., & Pischke, J-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Angrist, J. D., & Pischke, J-S. (2015). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

Ataneka, A., Kelcey, B., Dong, N., Bulus, M., & Bai, F. (2023). PowerUp R Shiny App (v. 0.9) Manual available at https://www.causalevaluation.org/uploads/7/3/3/6/73366257/r_shinnyapp_manual_0.9.pdf

Barcaui, A. (2025). ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. *Social Sciences & Humanities Open*, 12. https://doi.org/10.1016/j.ssaho.2025.102287

Bloom, H. S. (2005). *Learning more from social experiments: Evolving analytic approaches*. Russell Sage Foundation.

Bloom, H. S., Richburg-Hayes, L., & Black, A. (2007). Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis 29*, 30-59.

Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2022). PowerUpR: Power Analysis Tools for Multilevel Randomized Experiments. R package version 1.1.0. [Software]. https://CRAN.R-project.org/package=PowerUpR

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). Academic Press.

Diamond, J., Weiss, M.J., Hill, C., Slaughter, A., & Dai, S. (2025). MDRC's The Higher Education Randomized Controlled Trials Restricted Access File (THE-RCT RAF), United States, 2003-2024. (Version 5). [Data set]. Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR37932.v5

Dong, N., Kelcey, B., Maynard, R. & Spybrook, J. (2015). *PowerUp! Tool for power analysis*. Available at www.causalevaluation.org

Fitzpatrick, K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2), Article e19. https://doi.org/10.2196/mental.7785

Fulmer, R., Joerin, A., Gentile, B, Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health*, 5(4), Article e64. https://doi.org/10.2196/mental.9782

Glennerster, R., & Takavarasha, K. (2013). *Running randomized evaluations: A practical guide*. Princeton University Press.

Greenspan, R. L. (2025). Artificial intelligence policies in higher education: A randomized field experiment. *Journal of Criminal Justice Education*, 1–14. https://doi.org/10.1080/10511253.2025.2449603

Hedges, L., & Hedberg, E.C. (2014). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review, 37*(6): 445-489.

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025) AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15, Article 17458. https://doi.org/10.1038/s41598-025-97652-6

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*. *49*(4): 241-253.

Lambert, J., Martin, B. E., Stamm, R., White, S., & Kroger-Jarvis, M. (2026). Blinded but biased: Students prefer chatbot until they know it is one. *Journal of Nursing Education*, 0(0), 1–6. https://doi.org/10.3928/01484834-20260216-01

Mata, C., Russell, E., & Page, L. (2026). Scaling student support with conversational artificial intelligence. *EdWorkingPaper*: 26-1409. https://doi.org/10.26300/b0dp-sn77

Meyer, K., Page, L. C., Mata, C., Smith, E., Walsh, B.T., Fifield, C.L., Tyson, M., Eremionkhale, A., Evans, M., Frost, S., & Jung, E.E. (2024). Let's Chat: Leveraging Chatbot Outreach for Improved Course Performance. *EdWorkingPaper*: 22-564. DOI 10.26300/es6b-sm82

Murnane, R. J., & Willett, J. B. (2011). Methods matter: *Improving causal inference in educational and social science research.* Oxford University Press.

Nurshatayeva, A., Page, L.C., White, C.C., & Gehlbach, H. (2021) Are artificially intelligent conversational chatbots uniformly effective in reducing summer melt? Evidence from a randomized controlled trial. *Research in Higher Education*, 62, 392–402 (2021). https://doi.org/10.1007/s11162-021-09633-z

Page, L. C., Meyer, K., Lee, J., & Gehlbach, H. (2025). Conditions under which college students can be responsive to nudging. *Journal of Research on Educational Effectiveness.* 1–29. DOI 10.1080/19345747.2025.2481219.

Page, L. C., & Gehlbach, H. (2017). How an artificially intelligent virtual assistant helps students navigate the road to college. *AERA Open*, 3(4): 1–12. DOI: 10.1177/2332858417749220

Plewis, I., & Hurry, J. (1998). A multilevel perspective on the design and analysis of intervention studies. *Educational Research and Evaluation,* 4(1): 13-26. DOI 10.1076/edre.4.1.13.13014

Rhoads, C. H. (2011). The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics,* 36(1): 76-104. DOI 10.3102/1076998610379133

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Vol. 1195. Boston, MA: Houghton Mifflin.

Torgerson, D. J. (2001). Contamination in trials: is cluster randomisation the answer? *BMJ: British Medical Journal,* 322(7282): 355-357. DOI 10.1136/bmj.322.7282.355

Owan, V.J., Chukwu, C.O., Agama, V.U., Owan, T.J., Ogar, J.O., & Etorti, I.J. (2025). Acceptance and use of artificial intelligence for self-directed research learning among postgraduate students in Nigerian public universities. *Discover Education*, 4, Article 329. https://doi.org/10.1007/s44217-025-00770-6

Xu, S. & Ma, T. (2025). Depression intervention using AI chatbots with social cues: A randomized trial of effectiveness. *Journal of Affective Disorders*, 389. https://doi.org/10.1016/j.jad.2025.119760

Tables and Figures

Table 1. Mean characteristics of survey respondents and non-respondents

Variable	Non-respondent	Respondent	Difference	
Current chatbot exposure	0.48	0.54	-0.06	
	(0.50)	*(0.50)*		
Previous chatbot exposure	0.07	0.06	0.01	
	(0.25)	*(0.23)*		
Freshman	0.56	0.61	-0.05	
	(0.50)	*(0.49)*		
Repeating course	0.10	0.08	0.02	
	(0.30)	*(0.28)*		
High school GPA	3.70	3.66	0.04	
	(0.25)	*(0.37)*		
College GPA	2.81	3.36	-0.55	***
	(0.79)	*(0.59)*		
Honors	0.04	0.03	0.01	
	(0.20)	*(0.16)*		
SAT	1078.39	1122.73	-44.34	
	(161.06)	*(167.07)*		
Student's age	18.82	18.72	0.10	
	(1.52)	*(1.67)*		
Pell eligible	0.60	0.61	-0.01	
	(0.49)	*(0.49)*		
First-generation student	0.34	0.34	0.00	
	(0.48)	(0.48)		
Female	0.56	0.68	-0.12	
	(0.50)	*(0.47)*		
Asian	0.27	0.24	0.03	
	(0.45)	*(0.43)*		
Black	0.47	0.49	-0.02	
	(0.50)	*(0.50)*		
White	0.15	0.14	0.01	
	(0.36)	*(0.35)*		
Hispanic	0.18	0.17	0.01	
	(0.39)	(0.38)		
N	73	109		

Source: Authors' estimation.

Notes: Standard deviation in parentheses. Statistical significance: *p<0.05, **p<0.01, ***p<0.001

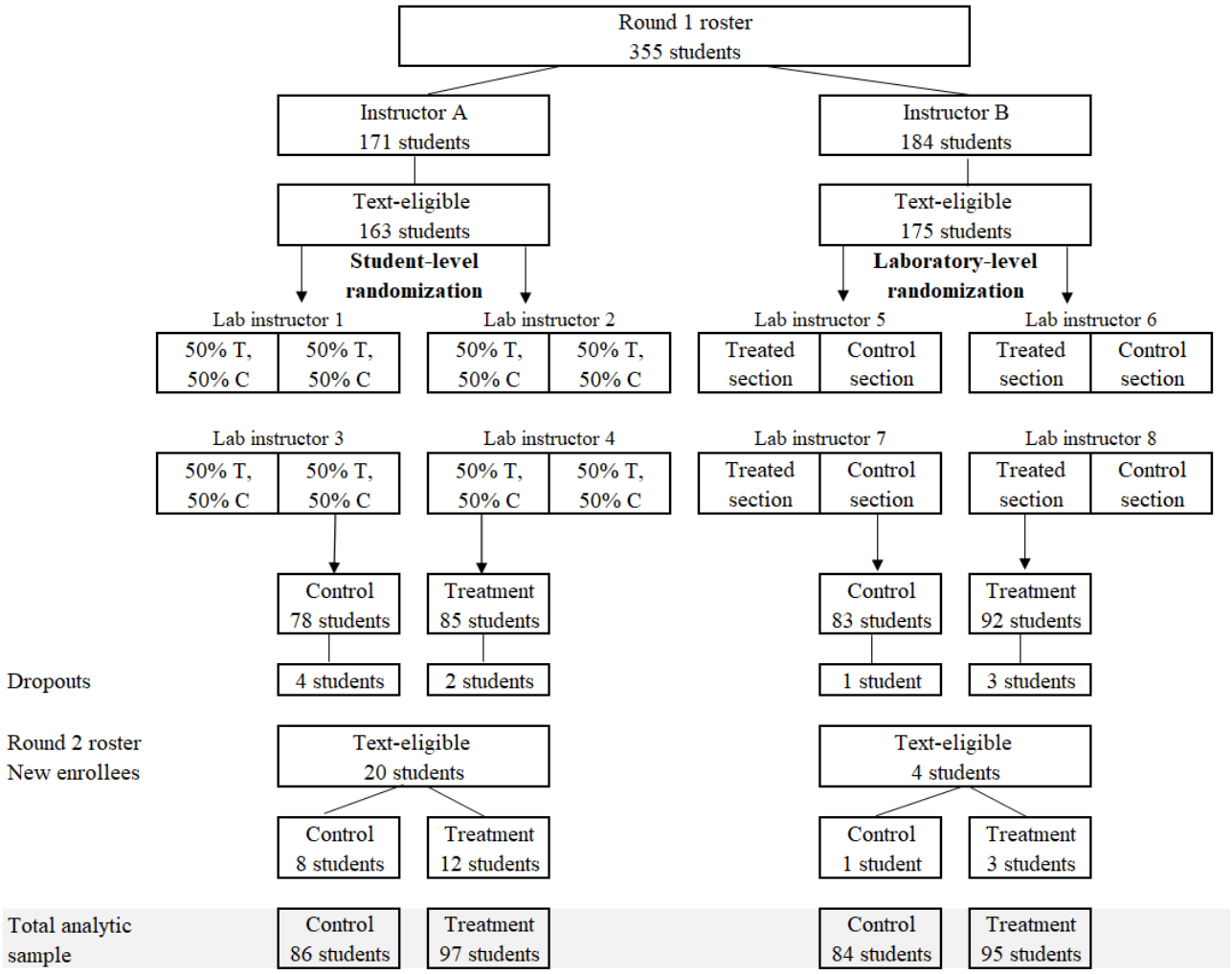
Table 2. Minimum Detectable Effect Size for Cluster Randomization at Different Levels

	Randomization level	
	(1) **laboratory**	(2) **individual**
Assumptions\Inputs		
Alpha Level (α)	0.05	0.05
Two-tailed or One-tailed Test?	2	2
Power (1-β)	0.80	0.80
Rho (ICC)	0.10	n.a.
P	0.50	0.50
R_1^2	0.20	0.20
R_2^2	0.02	n.a.
g*	2	2
n (Average Cluster Size)	25	25
J (Sample Size [# of Clusters])	64	64
Outputs		
M (Multiplier)	2.85	2.80
T_1 (Precision)	2.00	1.96
T_2 (Power)	0.85	0.84
MDES	**0.25**	**0.13**

Source: Authors' calculation using PowerUp! Version: 05/12/2019 [© Nianbo Dong and Rebecca A. Maynard].

Notes: n.a. stands for not applicable. We assume a total of 8 lecture sections, each with 8 associated laboratory sections, resulting in 64 laboratory sections overall. Each lecture enrolls approximately 200 students, with about 25 students per lab section. For reference, see Bloom et al. (2007) for typical R-square values of covariates at both the individual (level 1) and group (level 2) levels. For typical intraclass correlation coefficient (ICC) values, commonly ranging from 0.10 to 0.20, refer to Hedges & Hedberg (2014) and Bloom et al. (2007).

Figure 1. Analytic sample and enrollment across randomization approaches. Spring 2024



Source: Authors' elaboration.

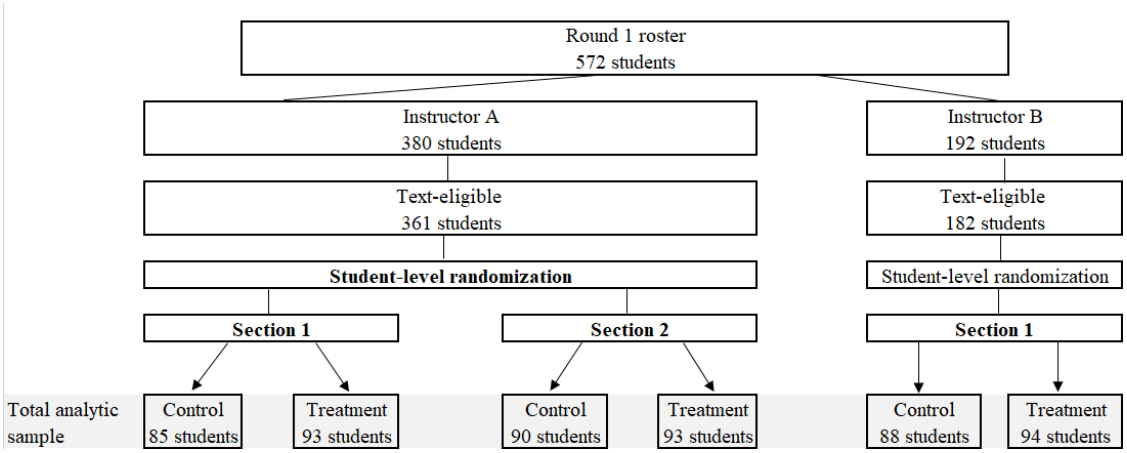
Notes: We implemented individual-level randomization at the lecture level, using blocks defined by lab sections to maintain a near 50/50 split between treatment (T) and control (C) within each lab section. For cluster-level randomization at the lab level, we randomized entire lab sections—assigning one section to treatment and one to control within each lab instructor's group. The randomization process was repeated to accommodate new enrollees. No students switched lecture or lab sections between the first and second rounds of randomization. We assessed covariate balance after each round of randomization.

Figure 2. Selected chatbot messages

Week 2 – Spring 2024			
Intro script launched to students who enrolled in the course during drop/add Hi name_first! I'm the chatbot for Principles of Chemistry I (CHEM 1211K) I'm working with your professor to help you stay on track. I'll send you course reminders and tips to succeed. You can text me questions anytime! Save my number, hit me up and I'll do my best to get you the answer. If you don't want these messages, just text #PAUSE to stop (but I hope you'll give me a chance).	**Weekly Digest:** Concepts 1.6-1.11 Dimensional Analysis, Problem solving strategy, Atom, Atomic Mass, Isotopes, Average Atomic Mass, Periodic Table, Ions Assignments: HW2 (due DATE by 12pm) Quiz 1: in class DATE CHEM 1211K WEEKLY DIGEST Hi name_first! You made it through week 1! This week we'll be working on atoms and revisiting our old friend the periodic table. Here's what's due this week: gastate.view.usg.edu/d2l/home/2989251 HW 2 (due 1/22) Lab Lecture + Lab Session 1 Quiz 1 (Friday in class) Labs start this week, attend both the lab lecture and the lab session. Your first quiz will be given in class on Friday. Make sure to study up! Syllabus link	**Targeted Support:** students who have not submitted HW for week 1 Hi name_first Looks like you might have missed your first HW assignment for CHEM 1211 Your best 6 (out of 8) count toward 10% of your grade. Don't worry – just make sure to get your HW in for this week! If you're having trouble submitting the HW make sure to reach out to Prof. Professor Name or if you need a bit of extra help check out the STEM tutoring lab STEM Tutoring	**Quizme:** Test students' understanding of key concepts presented in week 2 Ready for your Week 2 Chem check-in? ✅ Type #wk2check to check on the key concepts covered this week (don't worry, this check is not for a grade). There's a quiz tomorrow in class. If you want a refresher from last week's material #wk1check *Pro Tip – Quizzes make up 10% of your grade. Take it seriously, review concepts covered in class and in the HW. If you need extra help visit STEM Tutoring or chemistry.gsu.edu/ctc/ Don't forget your NON PROGRAMMABLE calculator.

Source: Georgia State University

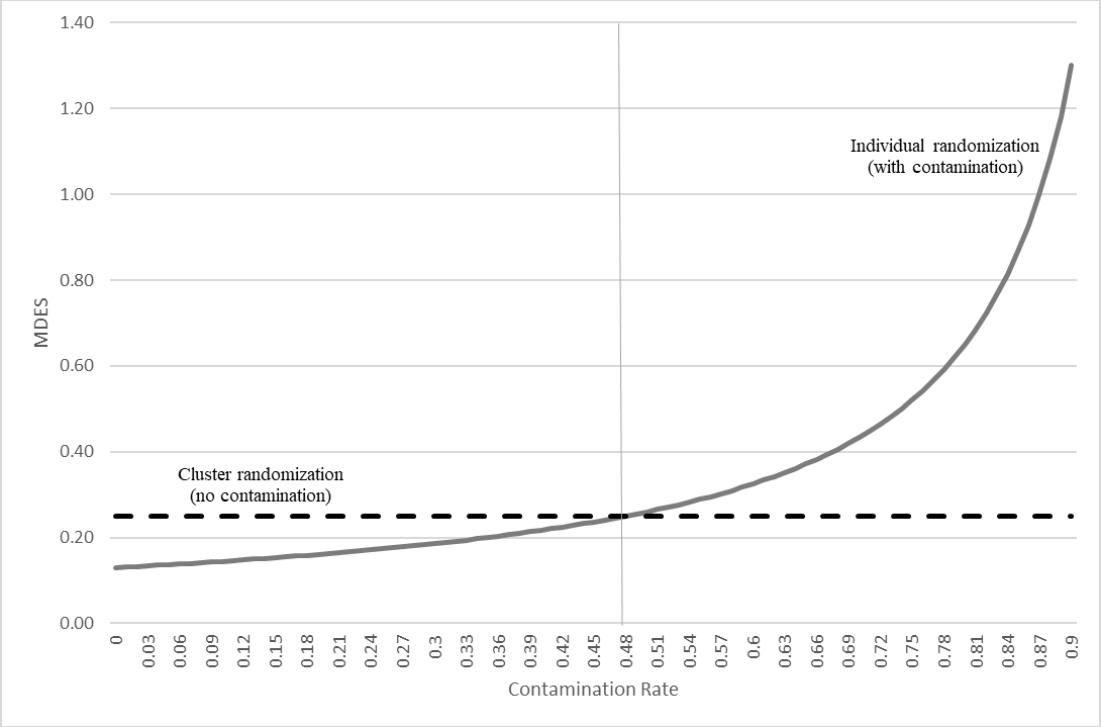
Figure 3. Analytic sample and enrollment in a second semester of randomization, Fall 2024.



Source: Authors' elaboration.

Notes: The balance between control and treatment groups across covariates was assessed after randomization.

Figure 4. Minimum detectable effect size by randomization level and contamination rate



Source: Authors' elaboration.