Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1083

Examining the Relationship Between Randomization Strategies and Control Group Crossover in Higher Education Interventions

Catherine Mata
Brown University

Katharine Meyer
Brookings Institution

Lindsay Page
Brown University

This article examines the risk of crossover contamination in individual-level randomization, a common concern in experimental research, in the context of a large-enrollment college course. While individual-level randomization is more efficient for assessing program effectiveness, it also increases the potential for control group students to cross over into the treatment group, thus biasing treatment effect estimates. This study provides empirical evidence from a pilot intervention in two sections of a college-level introductory chemistry course, where a course-specific chatbot was introduced. We tested two randomization strategies: simple student-level randomization and laboratory-level randomization. We hypothesized that the greatest risk for crossover would have occurred under the simple individual randomization approach, however, no crossover occurred in either condition. Survey responses and system usage data indicate that this was not due to a lack of interaction among students or disinterest in the chatbot. These findings suggest that student-level randomization, even in an in-person course setting, can proceed with minimal risk of contamination for testing our focal intervention.

VERSION: November 2024

Suggested citation: Mata, Catherine, Katharine Meyer, and Lindsay Page. (2024). Examining the Relationship Between Randomization Strategies and Control Group Crossover 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 Control Group Crossover in Higher Education Interventions

Catherine Mata
Brown University

Katharine Meyer
Brookings Institution

Lindsay Page*
Brown University

This version: November 5, 2024

Abstract

This article examines the risk of crossover contamination in individual-level randomization, a common concern in experimental research, in the context of a large-enrollment college course. While individual-level randomization is more efficient for assessing program effectiveness, it also increases the potential for control group students to cross over into the treatment group, thus biasing treatment effect estimates. This study provides empirical evidence from a pilot intervention in two sections of a college-level introductory chemistry course, where a course-specific chatbot was introduced. We tested two randomization strategies: simple student-level randomization and laboratory-level randomization. We hypothesized that the greatest risk for crossover would have occurred under the simple individual randomization approach, however, no crossover occurred in either condition. Survey responses and system usage data indicate that this was not due to a lack of interaction among students or disinterest in the chatbot. These findings suggest that student-level randomization, even in an in-person course setting, can proceed with minimal risk of contamination for testing our focal intervention.

JEL codes: C9, I23, D9
Keywords: randomization, causal inference, behavioral economics, nudge, 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 conference and the Southern Economic Association annual meeting for helpful feedback. Institutional support was provided by the Annenberg Institute at Brown University and Brookings Institution. All errors are our own.

Examining the Relationship Between Randomization Strategies and Control Group Crossover in Higher Education Interventions

Randomized controlled trials (RCT) are the method of reference in quantitative causal inference (Angrist & Pischke, 2009, 2015; Murnane & Willet, 2011). However, field experiments involve non-trivial design decisions. For instance, when implementing experiments in social settings like schools and classrooms where students are constantly interacting (Rhoads, 2011), 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 (Bloom, 2005; Shadish et al., 2002; Plewis & Hurry, 1998). Such contamination can attenuate the estimated treatment effect (Rhoads, 2011; Torgerson, 2001). Further, ethical concerns may stand in the way of employing an experimental design if it is considered unethical to deny students in the control group access to an intervention that may have positive benefits (Glennerster & Takavarasha, 2013).

Previous work has discussed the risk of contamination when implementing individual-level randomization in educational and other social settings (Rhoads, 2011; Bloom, 2005; Shadish et al., 2002; Plewis & Hurry, 1998). This contamination can take two different forms. First, control-group students can actively seek 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 positive effects of an intervention could spill over from the treated students to those assigned to 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. In this

article, we focus on measuring the extent of crossover contamination, with an exploratory examination of potential spillover contamination to better understand the viability of estimating the impacts of our focal intervention when implemented in a large-enrollment, in-person college course.

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 is still more powerful than cluster-level randomization with none. The key question, then, is how much contamination can be expected and tolerated in a given context.

Here, we build on Rhoads' (2011) foundational work to investigate this question in a higher education context. Specifically, we present empirical evidence of the level of crossover encountered in two different approaches to randomizing students to an intervention implemented in a large-enrollment, in-person, undergraduate course. The focal course is an introductory chemistry course in which students participate weekly in three one-hour-long, in-person lectures (attended by approximately 200 students) and a three-hour laboratory session (for which students are subdivided into groupings of up to 24). 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 campus. The data on which we report come from a first-semester pilot kicking off a multi-year experimental study of the chatbot tool specifically designed for the introductory chemistry course. In the study design process, we identified potential contamination as a key threat to the feasibility of the experimental study. Therefore, a primary goal of this pilot was to assess the level of contamination we might

expect and, in turn, determine whether student-level randomization was a viable option for evaluating the effects of course chatbots within in-person college courses.

Across two lecture sections, we piloted two different approaches to randomization: (1) simple student-level randomization of all students attending the same large lecture and (2) laboratory-level randomization, in which all students within a laboratory section were randomized either to treatment or control. For this study, IRB provided approval to waive informed consent. With both approaches, we conducted randomization among all students in the focal lectures who were eligible to receive text-message outreach according to university guidelines. Students assigned to treatment received text-based, course-specific outreach and communication throughout the semester. Students in the control group received business-as-usual course communication, none of which was via text messages. Course faculty did not mention the chatbot communication in class sessions or in the course syllabus. We describe the intervention in more detail below.

To address ethical considerations, any control group student who learned about and requested access to the course chatbot could opt into the chatbot communication, and any treatment group student receiving chatbot outreach could opt out at any time. In the first semester of implementation, our primary goal was to assess the extent of control group crossover (i.e., control students learning about and requesting access to the experimental course chatbot) and whether such crossover was more prevalent in the context of student-level compared to laboratory-level randomization.

Text-based outreach to treatment-assigned students began in the first week of classes and continued through the entire semester. To preview our findings, by the end of the semester, no control-assigned students requested access to the chatbot communication. About half of the students in the course were in their first year in college. In a large lecture setting, communication

and socializing may be challenging for students in this course. Nevertheless, students also spend three hours each week in their smaller laboratory groups. We were particularly surprised that, even with student-level randomization, no control-assigned students crossed over into the treatment condition. We conclude that we can continue to use student-level randomization in subsequent semesters of the study to assess the effect of the course-specific chatbot on student academic outcomes in this and other large-enrollment courses.

Intervention and Context

This 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 platform built by the technology company Mainstay and has been testing and expanding use of the tool in different aspects of student communication and support over time. GSU is a large, public university located in Atlanta, GA that serves a diverse student population. In our study sample, 66% of the students are female, 42% are Black, 15% are Hispanic, 61% are Pell-eligible, 24% are first-generation college-goers, 55% are freshmen, and 40% are continuing students who have previously completed one or more semesters at GSU.

Previous work (e.g., Meyer et al., 2023; Page et al., 2023; Page & Gehlbach, 2017) 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. Both of the courses in that research were offered online, so students in the courses seldom, if ever, interacted with each other in person. GSU is now testing the tool's effectiveness in Chemistry 1211 (CHEM1211), the university's first-semester chemistry course for STEM majors. Unlike the prior

courses in which the chatbot was tested, CHEM1211 is an in-person course. 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 during the pilot semester is pre-registered under Registry ID 18140.

In this article, we report on the first semester of implementation of the chemistry course chatbot in CHEM1211. Students enrolled in two sections of CHEM1211 in the Spring of 2024 were randomized either to control or treatment (see Figure 1). Across both lecture sections, students participate in large lectures of approximately 200 students each and small laboratory subsections of approximately 24 students each. Two experienced instructors, referred to here as Instructor A and Instructor B, each with extensive experience teaching CHEM1211, agreed to collaborate on the intervention. Students enrolled with Instructor A and who upon enrollment agreed to receive any university communication via text message (i.e., "text-eligible" students) were randomized at the student level. 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.

Text-eligible students enrolled with Instructor B were randomized at the laboratory level, with randomization nested by laboratory instructor. Students enrolled in the same course section are subdivided into eight laboratory sections, taught by four different laboratory instructors, each of whom oversees two laboratory sections. 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 in the treatment condition and 170 students in the control group, 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 outreach and support with specific course content and general academic competencies via text messages. 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 outreach. 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.

Data and Methods

We rely on three types and sources of data to assess messaging system engagement among treated students and intervention crossover among control students: (1) student-level administrative records held by the university, (2) student-level chatbot engagement metrics from the messaging platform, and (3) student-level 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 engagement with the chatbot: Using records from the Mainstay platform, we observe each outreach message sent from the chatbot to students and each response from students to the chatbot. This data enables 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 in to chatbot communication. Second, we can gauge student engagement by tracking the number of messages students send to the chatbot.

End-of-course survey: We conducted an end-of-course student survey tailored to 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 how they perceived the chatbot to have affected their course time management, content knowledge, and sense of chemistry belonging. Treatment group survey questions included

Likert-scale items through which students rated the usefulness of the chatbot and reported on whether and the extent to which they relied on it as a key channel of communication and/or source of course-specific information. From control group students, the survey was designed to gather insights into the reasons why control group students might have been more or less likely to request access to the chatbot. Students in the control group were asked 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.

In our analysis, we first estimate the proportion of control group students who opt-in for chatbot outreach within each randomization setting and compare the levels of crossover between the two randomization settings. We then contextualize the extent of crossover using results from an end-of-course survey.

Results

Treatment-assigned students received text-based communication throughout the entire academic semester. To our surprise, by the end of the semester, none of the control group students in either randomization scheme requested access to the course chatbot.

We investigated possible explanations for the lack of control group crossover. First, it is possible that students are 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. Only 5.7 percent of the treatment group students (11 students in total) opted out of receiving chatbot communications. All of the students

who opted out did so during the first month of chatbot outreach, and approximately half of these students dropped or withdrew from the class. Further, 69 percent of the treatment group students sent at least one message to the bot, and students who replied to the chatbot sent an average of 18 messages over the 17 weeks of the semester. One of the features available to students through the chatbot was the ability to request weekly quizzes, designed to provide practice for upcoming exams. About 41 percent of treated students, and 59 percent of those who responded to the chatbot, launched at least one quiz. Each quiz could be launched multiple times. On average, treated students launched 5.6 quizzes, with the number of launches per student ranging from 1 to 46. In summary, treatment students were highly engaged with the chatbot, as evidenced by both low opt-out rates and active participation. This indicates that lack of crossover is not due to lack of interest or receptivity.

We also considered students' self-reported perceptions of the course chatbot as indicated in their responses to the end-of-course survey. In prior instances of course chatbots, students generally reported enthusiasm for the chatbot, with over 90 percent of survey respondents recommending its continued use in the course and expansion to other courses at GSU (Meyer et al., 2023). As noted above, these prior studies were situated within courses taught online rather than in person. In in-person course contexts, students may feel less positively or perceive less of a need for chatbot communication. A limitation of the data on which we rely here is that procedural issues slowed our launch of the end-of-course survey, and we only received 12 responses in total. In subsequent semesters, we will be implementing the survey earlier, and faculty have agreed to incentivize survey completion with course extra credit. Nevertheless, of the treatment students who responded to this first survey, all reported reading all the course chatbot messages they received during the semester, and all but one found the messages somewhat or extremely helpful.

Finally, all treatment group respondents recommended continuing to use the chatbot in the same chemistry course in future semesters and expanding its use to other courses. These measures similarly suggest that treatment students find value in the course-based chatbot communication.

We also looked to the end-of-course survey as an additional source of evidence with which to gauge the extent of treatment contamination. Despite an elevated risk of contamination in an in-person course context, among control group survey respondents, none reported knowing about the existence of the chatbot implemented in their course. This is consistent with treatment students not discussing or passing the text content on to 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 limited survey data, both treatment and control students reported having connections with class peers. Almost all control group survey respondents indicated interacting with between one and six classmates outside of class meetings, and all control respondents indicated having another student they could ask questions about class material, if needed. As noted above, we recognize the limitations of the survey data, given the very low response rate.

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 24 students. In this context, we tested two different approaches to randomization: (1) simple 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 in to receiving the communication.

We hypothesized that the greatest risk for crossover would have occurred under the simple individual randomization approach, as this would lead to the scenario where both treatment- and control-assigned students were in small-group laboratory sections 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. 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.

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

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.

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

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

Meyer, K., Page, L. C., Mata, C., Smith, E., Walsh, B.T., Fifield, C.L., Eremionkhale, A., Evans, M., & Frost, S. (2023). 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.

Page, L. C., Meyer, K., Lee, J., & Gehlbach, H. (2023). Conditions under which college students can be responsive to nudging. *EdWorkingPaper*: 20-242. DOI 10.26300/vjfs-kv29

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

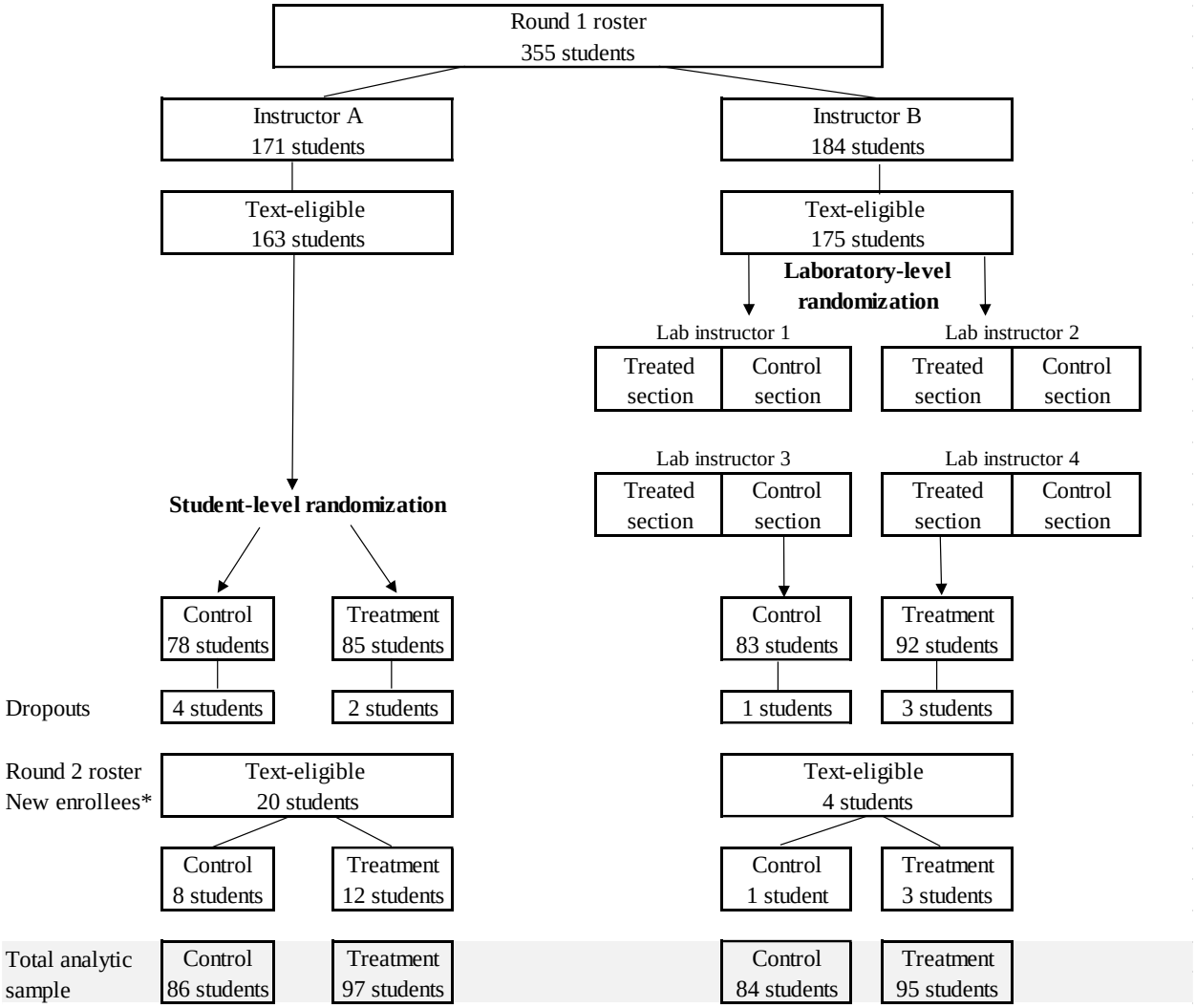
Rhoads, C. H. (2011). The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics,* 36(1): 76-104. DOI 10.3102/1076998610379133

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

Figures

Figure 1. Analytic sample and enrollment across randomization approaches



Source: Authors' elaboration.

Notes: * new enrollees were randomized too. No student was switching across lectures or changing laboratory sections between round 1 and round 2 of randomization. The balance between control and treatment groups across covariates was assessed after each randomization round.

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.