Lacking motivation or lacking productivity to explain failure: a field experiment

Daniel Dench[*]

August 30, 2023

Abstract

Why do some college students fail where others succeed? Is it for a lack of motivation in allocating time to work or a lack of productivity in time input? I run an experiment that substantially induces greater effort among students. The intervention, which includes counting homework towards students' final grade, increases attempting assignments by 72 percentage points, 76 percentage points for high GPA students, and 68 percentage points for low GPA students. Meanwhile, inducing students to attempt problems increases the likelihood they get similar problems on exams correct by 3.4 percentage points overall, 4.1 percentage points for low GPA students, and 2.9 percentage points for high GPA students. Students spend about an equal amount of time on assignments given an attempt. These results provide evidence that low GPA students are less responsive to effort-inducing interventions, while being equally productive learners. **Keywords:** nudge, incentive, motivation, online learning **JEL No.** I23, J24

[*] Georgia Institute of Technology, 221 Bobby Dodd Way NW, Atlanta, GA 30332, e-mail: dench@gatech.edu

[†] **Acknowledgement:** I would like to thank my adviser Ted Joyce for exceptional advice throughout this process. I would also like to thank Michael Grossman, Henry Saffer, and Dhaval Dave, three outstanding committee members who provided input. I would also like to thank seminar participants at Baruch College's Department of Economics and Finance, The CUNY Graduate Center's Economics Program, Baruch College's Marxe School of Public and Administration Affairs, The Georgia Institute of Technology's School of Economics, and West Virginia University's John Chambers College of Business and Economics. This trial was pre-registered on January 12, 2018 and the RCT ID was AEARCTR-0002669.

1 Introduction

Since 1990, the graduation rates for bachelor's degrees in the United States, conditional upon enrollment, have seen growth. Nevertheless, only about half of the enrolled students eventually graduate from college (Denning et al., 2021). Failure can substantially diminish the perceived likelihood of returns on post-graduate employment, leading many students to drop out (Stinebrickner and Stinebrickner, 2014). While the exact reasons for such failures are not fully understood, compelling evidence suggests that many of these reasons can be traced back to characteristics and skills students possessed upon entering college (Bowen et al., 2009) Yet, it remains ambiguous whether this variation arises from differences in the amount of productive time invested or the efficiency of that time input.

I conduct a field experiment to examine the impact of interventions designed to motivate students to complete homework assignments, and to see how this effort subsequently affects their exam performance. The primary intervention involves counting the homework toward students' grades, rather than merely suggesting a problem. To execute this, I randomized students into two groups. One group was assigned six out of 12 assignments that contribute to their final grade, while the second group received these as recommendations. For the other six assignments, the roles were reversed: the second group had these count toward their grade, while the first received them as recommendations. Thus, each assignment effectively became its own experiment, with roles switching between treated and control groups. By pooling the results, I obtained the average effect of these twelve individual experiments, even though they differ by a notable dimension. Specifically, four assignments contribute to 1 percent of the final grade for the assigned group, another four count for 2 percent, and the remaining four account for 3 percent. Consequently, the pooled effect represents the average of these varied incentives.

Separately, I assess the impact of informing students that a particular homework is likely to appear on the exam and how this affects their attempting the assignment. Unlike the previous intervention that works by randomized group assignment, this treatment applies at the assignment level. Students receive notifications via email that one of every two assignments might contain problems or similar problems that might feature in the exam. They account for six out of the total twelve assignments. Although this part of the study is observational rather than experimental, evidence suggests that this approach is significantly less effective than counting assignments toward the final grade. Moreover, this observation does not interfere with the core aspect of the primary experiment, which examines the differences in effort and efficiency of effort between high and low GPA students.

To test the effect of increased motivation on exam performance, 12 related problems to these assignments appear on students' midterms and finals. Since each homework acts as its own experiment, I recover estimates for the effect of attempting each homework assignment on each corresponding problem on the exam. I also pool these experiments to obtain the average effect of this experiment. The experimental setting comprises a set of reduced-lecture intro to microeconomics classes where students complete all of their homework in an online learning module.

I find that when students' homework counts towards their final grade, the likelihood of them attempting an assignment increases by 71.8 percentage points. In contrast, informing a student that a problem similar to their homework assignment is likely to appear on the exam raises the likelihood of attempting the assignment by only 3.0 percentage points. Additionally, when students attempt assignments—induced by the grading intervention—the probability that they answer related problems on the exam correctly increases by 3.4 percentage points. Elevating the weight of an assignment from one to two percent boosts the likelihood of attempting it by 4.6 percentage points. However, there is no noticeable increase in effort when the assignment's weight goes from 2 to 3 percent. Students with above-median GPAs show a stronger response in attempting assignments when the assignment counts toward their grade or when informed that a problem will appear on an exam. Yet, the effect of attempting an assignment on exam performance remains consistent across high and low GPA students.

This paper contributes to a large body of research on incentives. Some studies emphasize monetary incentives (Bettinger, 2012; Angrist and Lavy, 2009; Angrist et al., 2009; Le, 2015; Behrman et al., 2015), while others focus on grading incentives (Romer, 1993; Grodner and Rupp, 2013;

Grove and Wasserman, 2006; Trost and Salehi-Isfahani, 2012; Emerson and Mencken, 2011; Artés and Rahona, 2013). My paper focuses on the latter incentive literature but advances this literature in several key ways. First, I follow students through the educational process: from influencing their effort allocation to observing how that effort affects their outcomes. My experimental setup is similar to that of Trost and Salehi-Isfahani (2012). However, unlike Trost and Salehi-Isfahani (2012), I measure effort by tracking the time students engage with an assignment. This helps determine if variations between high and low GPA students stem from time spent or from productivity. This distinction matters, as efforts to enhance outcomes through increased effort depend on productivity of that effort. Moreover, this represents the largest experiment involving randomized grading as part of an intervention, allowing me to study effects on specific groups, such as high or low GPA students. This offers insights into whether top-performing students are simply more motivated or more efficient.

Additionally, I am the first to study grading treatment intensity. Understanding responses to grading intensity is vital, as courses have a fixed GPA credit, setting an upper limit for incentive allocation. If students respond mainly to extensive margins for grading, teachers might have more influence over students' overall effort in learning. However, if students react to grading on the intensive margins, teachers might prioritize key content. I also test a novel intervention, informing students if a homework problem will appear on an exam, both by itself and in conjunction with grading.

Lastly, this study serves as a potential guide for enhancing work in online or reduced lecture settings. An experiment by Bowen et al. (2014) finds that online courses yield similar test scores to conventional courses. However, other experiments (Joyce et al., 2015; Alpert et al., 2016) and an instrumental variable approach (Bettinger et al., 2017) find that reducing class time and increasing online classes can negatively impact performance. Reduced in-person interaction presents challenges in motivating students. This study aims to guide professors in optimizing effort interventions in increasingly online environments.

2 Framework

I begin with the premise that students aim for a good grade[1] (S), providing benefits at a rate of w. This rate encompasses all financial and non-financial rewards linked to a higher grade. I assume that three factors determine S: ability (A), classroom capital (K), and the time allocated to homework in the class (E). I further posit that students choose E, which may come at a cost and could be associated with A, to optimize the objective function

$$wS(E, K, A) - c(E, A). \tag{1}$$

Assignments that count towards a student's final grade are more likely to be completed by them. The more weight an assignment carries towards the final grade, the more likely students are to invest time in it, provided the effort remains constant. Furthermore, informing students that a problem similar to their assignment might show up on their exam reduces the unpredictability of their effort's payoff, thereby increasing the chances they will finish the assignment.[2]

3 Experimental setup

3.1 Setting

The study is conducted at a college, spanning eight large sections of principles of microeconomics during the Spring semester of 2018. These sections follow a reduced time format, featuring a single 75-minute lecture weekly, in contrast to the usual format of two 75-minute lectures per week. The mandated text is the online edition of "Principles of Economics" by Gregory Mankiw, accessible via Cengage's learning platform. This platform provides a digital version of the textbook, problem sets, flashcards, informational videos, and additional activities chosen by the professor from a

[1] I opt to use grades instead of human capital (H) as done in a similar model by Dee and Jacob (2012). The value of grading might reflect the human capital it represents or the signal it sends to potential employers. I don't specify which is more crucial since it doesn't change any predictions from my model.

[2] Refer to appendix A for additional information.

pre-set array provided by Cengage. For this class, the platform is set up to offer pre-lecture quizzes that correspond with the readings students should complete before the linked lecture. Post-lecture, students have the option to tackle a series of assignments: these are more demanding and time-intensive problem sets centering on fundamental concepts. Assignments are of two kinds: those that contribute to a student's final grade and those recommended but not factored into the final grade.

Before the semester starts, I randomly divide students into two groups, referred to as group A and group B. These groups are differentiated within each class. Students from both groups in every class receive an email containing a syllabus with directions on how to register on the online learning platform; the registration code they receive is group-specific. The platforms for both groups in each class are identical, except in experimental chapters where the treatment takes place. Apart from the treatment outlined below, all students within classes have equal access to resources and materials.[3]

In the six experimental chapters, there are a total of 12 assignments, two per chapter. For each chapter, group A students are given one assignment that counts towards their grade and another that is simply recommended. Conversely, for the same chapters, Group B students receive the same assignment that was recommended for group A as graded, and the one that was graded for group A as just recommended. Therefore, each assignment essentially serves as its own separate experiment, with each task being randomized as graded either for group A or Group B students.

The assignments are structured so that students receive immediate feedback upon submission, with guidance provided for incorrect answers. Each problem within an assignment allows for three attempts, with slight variations in the problem for each try. A student's score for the assignment is the average of all their attempts, and this score is presented to the student, whether the assignment is graded or merely recommended. Within the Cengage Learning Platform, labels I designed indicate if the graded assignments will contribute 1%, 2%, or 3% towards a student's final grade.

For each of the 12 assignments, it is predetermined whether it will be highlighted for potential inclusion on the final exam, which I refer to as being "nudged" or not. If an assignment is nudged, students who receive that assignment as graded are sent the following email for chapter 5 as

[3] Every student enrolled in the reduced time classes participates in this experiment. This is because I obtained a waiver of consent from the institutional review board. A detailed discussion on this topic is available in Appendix B.

an example: "We strongly recommend that you give the Chapter 5 Post-Lecture Quiz your full attention. A question or questions like these are likely to appear on the exam. Please note this problem set accounts for 1% toward your final grade." On the other hand, students who are nudged for recommended assignments receive this message: "We strongly recommend that you become comfortable with Chapter 5 Recommended Practice Problems. Although they do not count towards your grade, questions like these are likely to appear on the exam."

The Cengage platform captures two specific metrics related to a student's effort on an assignment. Firstly, it tracks if a student has opened an assignment, which I define as an "attempt". Secondly, the platform records the duration a student spends on an assignment. This duration is the time elapsed between when a student initiates the assignment and submits it, assuming they don't close the browser window during that period.

3.2 Identification of nudging and grading on attempting an assignment

I compare the impact of information nudges with that of grading an assignment, a more conventional method to motivate student effort. Both grading and nudging serve as unique strategies. For clarity, envision the two-by-two matrix presented in Table 1. This table displays the treatment assignments for two representative assignments, labeled 1 and 7, designated for group A and group B. Mirroring the classic difference-in-difference layout, I have four sections, with the groups labeled on the top and assignments (akin to a time frame) on the left.

Cell a indicates the probability that students try an assignment when it is graded, and they receive a nudge about it. Cell b represents the likelihood of students attempting an assignment when it's merely recommended, but they still get a nudge. The difference between these probabilities, $(a - b)$, measures the influence of grading an assignment on the propensity of students to attempt it. On the other hand, the difference $(c - d)$ contrasts the likelihood of attempting an assignment for those assignments that are graded against those recommended, but without any hint of its importance for the exam. Both these differential measures offer unbiased estimates, given the random assignment of students to groups A and B.

3.3 Varying the grading intervention

I also investigate if there's a relationship between the grading percentage and student response, essentially, a dose-response to grading. The percentage of the homework grade, as it contributes to the student's final grade, varies. For instance, for Group A, the assignment from chapter 1 is worth 1 percent of their final grade, while for Group B, it's worth zero. In chapter 2, assignment 3 is worth 2 percent of the final grade for Group A and again, zero for Group B. By chapter 3, assignment 5 counts for 3 percent of Group A students' final grades. On the flip side, Group B gets assignments from chapters 1, 2, and 3 that respectively contribute 1, 2, and 3 percentage points to their final grades. This allocation continues similarly for chapters 4 through 6. For a detailed breakdown of these treatments, refer to table A.2 in appendix C.

3.4 The effect on the exam

On every professor's midterm exam, I include six multiple-choice questions from the first three experimental chapters of the course. These questions are directly linked to the concepts covered in the six graded and suggested assignments. The final exam, which is consistent across all classes, adds another batch of six multiple-choice questions stemming from the latter three experimental chapters of the semester. The degree to which these exam questions mirror the assignments varies, allowing me to gauge the range of understanding, from memorization to a deeper comprehension of the concepts. For a closer look at the questions used in the exams, refer to Appendix D. Additionally, table A3 provides insights into the variations between the practice problems in the assignments and those posed in the exams.

Due to the randomization of grading assignments between groups, I leverage this external treatment to gauge the impact of attempting an assignment on the likelihood of answering a related question correctly in the exam. I anticipate a considerable and significant influence of grading on a student's decision to attempt an assignment. Given that each assignment serves as its own grading experiment, I can obtain both the reduced form and instrumental variable estimates to assess the effect of attempting an assignment on performance in related exam questions, considering variations across assignments.

3.5 Sample assignment and attrition

Figure 1 illustrates the sample division and its progression throughout the semester. It's important to point out that seven students from group A and seven from group B mistakenly registered for the other group's online module. This error likely happened because these students obtained a syllabus from a peer in the opposite group. I classify them based on the group they registered for, not the syllabus they were given. Given that these students represent only 1.6% of the total sample, their inclusion or exclusion doesn't significantly affect the results in terms of magnitude, direction, or statistical significance. Even if I were to classify them based on the syllabus they received, the results remain consistent. From the original 833 students, 738 took the final exam. The dropout rates are very close between the groups: 12.2% for group A and 12.7% for group B.

4 Empirical specification

4.1 Time spent specifications

One way to think about this experiment is simply as a number of pooled experiments across assignments. In order to pool experiments to get the average effect across experiments, I specify a model such that

$$a_{it} = \alpha_0 + \alpha_1 G_{it} + \alpha_2 N_t + \alpha_3 G_{it} \times N_t + \omega_t + \nu_i + \eta_{it}. \tag{2}$$

a_{it} represents the probability that student i attempts assignment t. G_{it} equals one if an assignment is graded and zero otherwise. N_t equals one if an assignment is nudged and zero otherwise. Since nudges occur both for graded assignments and recommended assignment, I include an interaction $N_t \times G_{it}$. $\alpha_2 + \alpha_3$ tells you how nudging a graded assignment increases the probability of an attempt. Students receive one nudged assignment per chapter. I include chapter fixed effects, ω_t and student fixed effects, ν_i, to absorb residual variation. Both sets of fixed effects are uncorrelated with the three variables of interest, since grading is randomized and each student receives a nudge for at least one of their assignments within each chapter.

Additionally, I estimate models that consider the time spent on an assignment and the score, conditional on an attempt. Due to variations in time spent and score, contingent on assignment difficulty, I cannot definitively state that the estimates of the effect of nudging on these metrics are unbiased. However, these models could provide descriptive insights, and I thus include them in the analysis.

Furthermore, I estimate models where I omit the nudging and the interaction term to gauge the effect of grading, but I use assignment fixed effects instead of chapter fixed effects. Nudging cannot be integrated into these models since assignments are either nudged or not. Finally, in other models, I look at dose response models by replacing G_{it} with three dummy variables that indicate whether the assignment is worth 1%, 2% or 3% towards a students final grade.

4.2 The effects of grading and attempting assignments on exams

I specify a reduced form model as

$$c_{it} = \beta_0 + \beta_1 G_{it} + \psi_t + \upsilon_i + \epsilon_{it}. \tag{3}$$

c_{it} represents the probability of correctly answering problems related to assignment t on the exam. While assignment and person fixed effects absorb residual variation, they aren't crucial for generating unbiased grading estimates. To gauge the effect of attempting an assignment on correctly answering its related problems on an exam, a single valid instrument is necessary. If grading effects in the first-stage are sufficiently pronounced, this allows for estimating the effect of a_{it} on exam question correctness. Heterogeneous effects of attempts across assignments can also be determined, providing qualitative insights into intervention variations across concepts and the extent of assignment modifications from homework to exams.[4]

[4] Estimating the effect of nudges on exam correctness for problems related to nudging assignments is not feasible.

5 Results

5.1 Balance

Student characteristics across groups are presented in Table 2. The randomization process results in closely matched characteristics between groups. Although Group B have GPAs that are marginally significantly different from Group A, this difference is only approximately 1/7 of a standard deviation (S.D.) in GPA within my sample. It's noteworthy that the crossover design empowers me to implement student fixed effects, thereby minimizing the impact of any random characteristic variations on the study's findings. A joint chi-square test comparing assignments to either Group A or Group B based on student characteristics demonstrates balance: $p = 0.24$ at the semester's commencement and $p = 0.21$ at its conclusion.

5.2 Effect of grading and nudging on attempts

The first-stage results in Table 3, column 1, shows the effects of the two treatments using student and chapter fixed effects. Grading an assignment increases the probability that it will be attempted when not nudged by 71.8 percentage points with a standard error of 1.3 percentage points. By comparison, telling someone an assignment like the recommended assignment is likely to be on the exam increases the probability that a student attempts it by 3.0 percentage points with a standard error of 0.8 percentage points. Telling them the same about their graded assignments increases the probability of an attempt by 1.8 percentage points with a standard error of 0.7 percentage points.[5] These results are not different if I remove chapter and student fixed effects, as expected. Excluding nudges and the interaction but including assignment fixed effects barely alters the effect of grading on attempts.

The imbalance in assignment difficulty between nudged and non-nudged assignments complicates this. Nudging, being assignment-specific, might bias results concerning exam correctness. A larger set of assignments could have facilitated random nudge assignments, ensuring balanced difficulty. Alternatively, an extra randomization dimension might have been helpful. Avoiding randomization between nudged and non-nudged groups was a conscious decision to retain power for detecting reduced form effects in the second-stage, as sample-splitting could compromise this ability.

[5] This is the estimated linear combination of the nudge and the interaction between nudging and grading.

Moving to models where I test how varying the percentage towards a student's final grades affects whether they attempt, in columns 3 and 4, there is no clear dose response. Assignments that were worth two to three percent towards students' final grade made the treated three to four percentage points more likely to attempt an assignment then when it was worth one percent. There was no significant difference between two and three percent toward students' final grade. In order to find a more effective margin for grading, I would have had to vary the worth of an assignment to be somewhere between zero and one.

Grading an assignment also means that given an attempt, they spend about 31.4 additional minutes on all attempts on that assignment than they otherwise would have. For their additional effort, the score they receive increases by 48.8 points out of 100 maximum. By contrast, nudging an assignment leads to only 3.5 more minutes spent for recommended assignments and 2.9 additional minutes for graded assignments. The effect on score decreases by 2.9 points when nudged.[6] This shows that ungraded assignments, even when students are told they are likely to be on the exam, are not treated as seriously as when they are graded.

5.3 Effects on the exam

The reduced form effects are detailed in Table 4. Employing the same independent variables as in column 2 of Table 3, the reduced form effect of grading an assignment on correctly answering related exam problems is 2.6 percentage points above a baseline of 53 percent. This impact remains consistent, irrespective of the inclusion of student fixed effects. The respective effects amount to 1.3 percentage points for the midterm and 4.0 percentage points for the final. Transitioning to instrumental variable estimation, the act of attempting an assignment enhances performance by 3.4 percentage points. For the midterm, this boost is 1.7 percentage points, while it rises to 5.1

[6] The effect of nudging on time spent and score should be interpreted with caution since the score depends on assignment difficulty and I did not balance assignment difficulty between assignments that would be nudged and assignments that would not be nudged at the beginning of the semester. Since these effects are identified between students across assignments, differences in assignment difficulty could bias the results. The same does not apply to assignment attempts since students do not know the difficulty of a assignment before attempting.

percentage points for the final.[7]

Returning to the nudge results, it's intriguing to speculate on the reduced form effects of nudging, particularly given the marginal adjustments observed in assignment attempts due to a nudge. For the subsequent calculations, I'll posit that the influence of an attempt prompted by grading mirrors that of a nudge-induced attempt concerning its bearing on correctly answering an exam problem.[8] Employing the first-stage nudge effects and the IV estimation correlating assignment attempts to correct answers in the exam, a rudimentary wald estimator can be derived.[9] By multiplying the 3.4 percentage point effect (of attempts on correct exam answers) with the 3.0 percentage point elevation in the likelihood of attempting a problem when it's nudged yet ungraded, I arrive at a reduced form estimate pegging the nudge effect at a subtle 0.1 percentage points.

5.4 Differences in motivation and productivity by prior GPA

The biggest differential response is between those with greater than or less than the median GPA as shown in Table 5. Those with greater than average GPA are more responsive to both grading and nudging an assignment. While both groups have large responses to grading, those with greater than a median GPA had slightly larger response than those with less than a median GPA. The response to nudging, however, is almost zero for those with less than the median GPA and double the overall effect for those with greater than the median GPA. This makes some intuitive sense. The greater expected learning is from time input, the greater the expected value of that time input. For example, if a student believes that no matter the effort they put in, it will not help them on the exam, the knowledge that a question will be on the exam will induce no extra effort.

[7] One plausible rationale for this disparity between midterm and final results could be students' intensified efforts post-midterm, aiming to compensate and hence, treating assignments with greater seriousness. However, no discernible shifts are observed in reactions to grading, nudging, or overall effort pre and post-midterm.

[8] This assumption perhaps overestimates the influence of nudges, considering the study time surge is roughly 8.9 times more pronounced for grading than nudging, conditional upon an attempt.

[9] The wald estimator is symbolized as $\hat{\beta} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}$, wherein $\hat{\beta}$ represents the projected parameter of a potential endogenous factor like attempting. The numerator to the right denotes the variation between the outcome variables of treated and control groups (for instance, correctly answering an exam question), whereas the denominator elucidates the discrepancy between the averages of the treated and control groups' endogenous variable (like attempting an assignment) (Cameron and Trivedi, 2005).

The split between GPA, however, did not predict how much students learned from attempting an assignment. In Table 6, the probability increase for low GPA students, 4.1 percentage points, was greater than for high GPA students, 2.9 percentage points, but it is not a significant difference. In addition, students with greater than median GPA spend about the same amount of time on assignments as students that have lower than median GPA, 39 and 36 minutes respectively. This suggests that with equal effort, students with lower than median GPA have about the same return on attempting an assignment as high GPA students.[10]

Perhaps the grading intervention has more effect in time input on low GPA students given an attempt. If this were the case the change in time input given an attempt would be greater for low GPA students than high GPA student. These differential effects by student can be observed in table 7 where I show the effect of grading, nudging and the interaction by student characteristics. The key result is that given an attempt, high GPA students increase time input a bit, although not significantly so, more than low GPA students. The non-differential response in output on the exam can not therefore be explained by differences in time response.

Turning to table 8, I look at the effect of grading on score received in a homework attempt. Here I show that that lower GPA students have a significantly higher increase in score given an attempt in comparison to higher GPA students. This is plausible evidence of more productive learning induced by grading in comparison to high GPA students. One possible reason for this could be that high GPA students start from a higher base score level, 43.3 on ungraded attempts and therefore the increase in score given an attempt induced by grading is a result of starting from a lower level of base knowledge and having more to learn given equal efforts on a given assignment.

The effects of attempting on answering exam questions correctly are treatment effects on compliers, those induced to attempt the assignment by grading, assuming there are no defiers (Angrist et al., 1996) and not average effects. The treatment effect differences therefore could be

[10] The response to both attempting and correctly answering problems is the same in direction when you split by below median and above median SAT scores. Taking SAT is not a necessary requirement for getting into this college and therefore is more missing than prior semesters' GPA. Prior semesters' GPA is less missing than perhaps at other colleges, because this introductory course is mostly restricted to students after their freshman year and few students matriculate to this college between the fall and spring semesters.

explained by differences in who is a complier between those attempting the assignment and those not. Across GPA you may expect this group to vary in their selectivity and that to be an explanation for why high GPA and low GPA students are equally productive in their efforts.

To examine this I will assume that those in the higher than median GPA category who are compliers were half as productive learners as those that were never or always takers. Since the majority of non-compliers in this group were always-takers (16.9%) rather than never-takers (7.4%) this is a not too unreasonable assumption. I will also assume that the compliers in the lower than median GPA group are twice as productive compared to those that were never or always takers. Again, not an unreasonable assumption given the never-takers (23.8%) outnumber the always-takers (9.2%). To calculate the average effects under these assumptions, I simply sum the effects observed in the complier groups multiplied by their percent of total and the assumed effects in the non-complier groups multiplied by their percent of the total. For the higher than median GPA students this is 75.7% multiplied by 2.9 percentage points plus 24.3% multiplied by 5.8 percentage points. For the lower than median GPA students this is 68.0% multiplied by 4.1 percentage points plus 32% multiplied by 2.1 percentage points. This yields an average effect of attempts of 3.6 percentage points for higher than median GPA students and an average effect of attempts of 3.4 percentage points for lower than median GPA students. That is to say, under these assumptions, the direction of the point estimate of average effects could flip the sign of the difference between these groups to those observed in compliers but still only be 6% different from one another.

One relevant question is how inducing more effort for low GPA students may impact them relative to high GPA students overall. For this exercise, I estimate the effect of below median GPA students putting in equal amounts of effort to high GPA students under the assumption that the average effects are the same as compliers. I will be conservative and assume that the differences in the effect of attempts on GPA that were observed are due to random chance. I assume that below median GPA students increase the probability they get a problem correct by 3.4 percentage points with attempts, the mean overall. I focus on graded assignments since both types of students tend to put a lot more effort in these assignments than other assignments.

For problems on the test for which above median GPA and below median GPA students have graded assignments, they answer these questions correctly 61.5 percent and 49.0 percent respectively. In addition, high GPA and low GPA students attempt graded assignments at a rate of 92.9 percent and 77.2 percent respectively. If low GPA students attempt assignments at the same rate as high GPA students, they will decrease the gap in score by 0.5 percentage points, 3.4×15.6. This accounts for approximately four percent of the gap in the percent correct between high and low GPA students.

While small, this estimate is conservative for the difference in the gap between high and low GPA students that can be made up by increased effort. I only account for the difference in effort on a single, albeit very related assignment. High GPA students also exert more effort in other class activities, such as attendance, readings, and pre-class reading quizzes.

Up until now, I've been assuming that the only thing that affects these 12 problems on the exam are attempts of 12 related assignments. Relaxing this assumption might help explain why above median GPA students are affected less by these experimental assignments than below median GPA students. In fact, high GPA students exert about 40 percent more total effort on online activities outside the 12 assignments from this experiment.

5.5 Variation along other characteristics

In Table 5, distinctions in reactions to grading or nudging an assignment based on baseline attributes are evident, particularly when considering assignment attempts. Firstly, notable heterogeneity by race emerges. Specifically, Asian and White students have a larger response to grading in comparison to their counterparts from other racial backgrounds. Furthermore, while White students seemingly benefit more from undertaking these practice assignments than either Asian or students from other races, the difference is not statistically significant. Secondly, in terms of gender dynamics, men and women demonstrate a roughly equivalent inclination towards grading. However, women appear to be more responsive to nudges. This larger response among women is further underscored when examining the effects of assignment attempts on related exam problems: women exhibit a more pronounced positive impact of attempting a problem than do men.

5.6 Problem specific heterogeneity

In investigating the varying impacts across different exams, I delve into problem-specific variations as shown in Figure 2. The figure shows the percentage of correct responses on the exam for the ungraded group on the x-axis against the effect of attempting the corresponding assignment on the y-axis. The overall picture presents a lack of discernible patterns tying problem difficulty to effects on the exam. Although a tentative negative correlation between ungraded student performance and the effect of assignment attempts emerges, it is riddled with outliers, particularly at the upper difficulty spectrum. My analysis centers on two specific problems, problem 1 and 12, that diverge notably from the mean effect across all problems.

Problem 1 pertains to shifts in supply and total expenditure. In their practice task, students read a described negative supply shift and its impact on the market's total revenue. While the practice leverages an elastic demand curve, the exam version employs an inelastic one. With a only 35.7% of the ungraded group providing the correct answer, it's evident the exam question proved challenging. Most students accurately identified the leftward supply curve shift but mostly erred in assuming a revenue decline. Their familiarity with the elastic demand curve from the practice might have inadvertently entrenched this sign. Thus, students who attempted the practice (because of grading) were 6.4 percentage points less likely to get the correct answer compared to their non-attempting peers.

In contrast, Problem 12 delves into the intricacies of cartel agreement collapses within the game theory framework. Although its premise mirrors the practice, the posed question diverges slightly. This proved daunting for many, with error rates among the ungraded students surpassing what might be expected from random guessing. This multi-step problem necessitates a clear and sequential thought process: what ensues when a two-party cartel jettisons its balanced output agreement, leading one party to increase its output? While the arithmetic is simple, the underlying logic is not. Without prior engagement through the practice, most fail. Yet, those who grappled with a similar question during their practice received a 16-percentage-point boost in their chances of getting the correct answer.

Thinking about the possible influence of external knowledge—potentially tethered to prior GPA—on the exam's effect, I can turn to Figure 3. This figure delineates the impact on each problem for students segmented by their standing with respect to the median GPA. A striking observation is the non-uniform advantage garnered by the below-median GPA students from these practice assignments. Indeed, a majority of the effects confidence intervals overlap.

In Problem 12, the dismal performance of the ungraded group—akin to random guessing—underscores the lack of any external beneficial knowledge to tackle this exam problem without the practice problem. Both high and low GPA groups benefit similarly from their engagement with the problem. In contrast, Problem 6 paints a different narrative. Here, the lion's share of students—whether they engaged with the related assignment or not—generally gets it right, suggesting an familiarity with the problem due to effort outside the practice problems. In this backdrop, those below the median GPA reaped more benefits from the practice assignment. This evidence supports the idea that the performance chasm between the two GPA cohorts stems more from effort disparity than from efficacy of effort.

6 Conclusion

In this research, I highlight the impacts of interventions on online college coursework effort and exam retention based on student GPA. While my findings confirm that grading boosts effort across the board, it is noticeably less effective for students with lower GPAs. Interestingly, even a direct nudge—such as informing students of the likely inclusion of a particular assignment question in the exam—barely impacts low GPA students and only marginally influences those with higher GPAs. This aligns with growing evidence suggesting that interventions with minimal direct incentives often yield limited effects at scale (DellaVigna and Linos, 2020).

Several hypotheses might explain why high GPA students, though more responsive to grading and nudging, don't demonstrate superior benefits from attempting practice assignments. One possibility is that both high and low GPA students have similar learning efficiencies with respect to time investment. Furthermore, having already invested significantly across the course, high GPA students might be on the diminishing returns portion of their learning curve, making additional efforts less fruitful. These ideas resonate with the patterns I observed in Figure 3.

Encouraging more effort from low GPA students could bridge the learning gap in online platforms. However, my findings indicate that broad interventions might inadvertently boost effort primarily among those already showing substantial baseline effort. A more tailored approach might be the answer. For instance, inspired by Dobkin et al. (2010), I suggest increasing interactions for under-performing students by mandating more participation. In the online context, this could translate to more graded assignments. Given that students seem largely unaffected by the weight of assignments toward their final grades, such an approach could spur increased effort.

Moreover, my research underscores potential limitations of online teaching tools like instant feedback on answers, particularly when foundational assumptions shift between assignments and exams. I must emphasize, however, that my conclusions are based on data from just 12 distinct assignments and their corresponding exam questions. A more extensive data set could provide deeper insights.

Finally, I demonstrate the potential of a crossover experimental design, centered around grading. The online nature of my experiment lends itself to replication across institutions. Through iterative adaptations of assignments, teachers could compile a rich repository optimized for learning. Using similar experimental designs, other facets of online learning could also be explored. I'm particularly intrigued by the potential of video-assisted learning punctuated by questions to engage students actively. Such a method might emulate and even surpass traditional classroom dynamics, promoting more intuitive understanding.

7 Declaration of Generative AI and AI-assisted technologies in the writing process'

Statement: During the preparation of this work the author used GPT 4.0 in order to edit grammar and readability. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

Alpert, W. T., Couch, K. A., and Harmon, O. R. (2016). A randomized assessment of online learning. *American Economic Review*, 106(5):378–82.

Angrist, J., Lang, D., and Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 1(1):136–63.

Angrist, J. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American economic review*, 99(4):1384–1414.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Artés, J. and Rahona, M. (2013). Experimental evidence on the effect of grading incentives on student learning in spain. *The Journal of Economic Education*, 44(1):32–46.

Behrman, J. R., Parker, S. W., Todd, P. E., and Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results from a social experiment in mexican high schools. *Journal of Political Economy*, 123(2):325–364.

Bettinger, E. P. (2012). Paying to learn: The effect of financial incentives on elementary school test scores. *Review of Economics and Statistics*, 94(3):686–698.

Bettinger, E. P., Fox, L., Loeb, S., and Taylor, E. S. (2017). Virtual classrooms: How online college courses affect student success. *American Economic Review*, 107(9):2855–75.

Bowen, W. G., Chingos, M. M., Lack, K. A., and Nygren, T. I. (2014). Interactive learning online at public universities: Evidence from a six-campus randomized trial. *Journal of Policy Analysis and Management*, 33(1):94–111.

Bowen, W. G., Chingos, M. M., and McPherson, M. S. (2009). *Crossing the finish line: Completing college at America's public universities*, volume 52. Princeton University Press.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Dee, T. S. and Jacob, B. A. (2012). Rational ignorance in education a field experiment in student plagiarism. *Journal of Human Resources*, 47(2):397–434.

DellaVigna, S. and Linos, E. (2020). Rcts to scale: Comprehensive evidence from two nudge units. Technical report, National Bureau of Economic Research.

Denning, J. T., Eide, E. R., Mumford, K., Patterson, R. W., and Warnick, M. (2021). Why have college completion rates increased? an analysis of rising grades. Technical report, National Bureau of Economic Research.

Dobkin, C., Gil, R., and Marion, J. (2010). Skipping class in college and exam performance: Evidence from a regression discontinuity classroom experiment. *Economics of Education Review*, 29(4):566–575.

Emerson, T. L. and Mencken, K. D. (2011). Homework: To require or not? online graded homework and student achievement. *Perspectives on Economic Education Research*, 7(1):20–42.

Grodner, A. and Rupp, N. G. (2013). The role of homework in student learning outcomes: Evidence from a field experiment. *The Journal of Economic Education*, 44(2):93–109.

Grove, W. A. and Wasserman, T. (2006). Incentives and student learning: A natural experiment with economics problem sets. *American Economic Review*, 96(2):447–452.

Joyce, T., Crockett, S., Jaeger, D. A., Altindag, O., and O'Connell, S. D. (2015). Does classroom time matter? *Economics of Education Review*, 46:64–77.

Le, V.-N. (2015). Should students be paid for achievement? a review of the impact of monetary incentives on test performance. *NORC at the University of Chicago*.

Romer, D. (1993). Do students go to class? should they? *Journal of Economic Perspectives*, 7(3):167–174.

Stinebrickner, R. and Stinebrickner, T. (2014). Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics*, 32(3):601–644.

Trost, S. and Salehi-Isfahani, D. (2012). The effect of homework on exam performance: Experimental results from principles of economics. *Southern Economic Journal*, 79(1):224–242.

Table 1: Example of identification on the effect of nudging[1] and grading on attempting an assignment

	Group A	**Group B**	**Identified across randomized groups**
Chapter 1 assignment 1	a=P(Attempt | graded and nudged)	b=P(Attempt | rec. and nudged)[2]	a-b : Effect of grading given a nudge
Chapter 4 assignment 7	c=P(Attempt | graded and not nudged)	d=P(Attempt | rec. and not nudged)	c-d: Effect of grading given no nudge
Identified across assignment	a-c: Effect of a nudge on graded assignment	b-d: Effect of a nudge on rec. assignment	

[1] Nudge refers to telling a student an problem like those contained in their graded or recommended assignment is likely to be on the test.

[2] Rec. is an abbreviation for recommended. Assignments that were recommended did not count towards students final grade.

Table 2: Characteristics by randomization group[1]

	Randomized			Completed Final		
	Group A	Group B	|Diff/S.D.|	Group A	Group B	|Diff/S.D.|
Prior School Performance						
GPA	3.1	3.2	0.12	3.1	3.2	0.16
SAT Verbal	550.1	549.2	0.01	550.7	550.7	0.00
SAT Math	611.4	610.6	0.01	614.9	617.4	0.03
School Experience						
Cumulative Credits	17.7	18.5	0.03	16.3	17.8	0.05
Underclass (%)	82.9	85.4	0.07	84.9	86.1	0.03
Attends Part Time (%)	4.2	2.9	0.07	3.1	2.2	0.06
Demographics						
Age	20.5	20.7	0.06	20.3	20.5	0.07
Female (%)	35.8	31.4	0.09	37.2	33.0	0.09
Speaks English at home (%)	13.7	14.5	0.02	15.4	15.5	0.00
Race/Ethnicity						
Asian (%)	53.2	50.0	0.06	54.9	51.5	0.07
Black (%)	7.3	8.0	0.03	7.2	7.5	0.01
Hispanic (%)	15.6	16.8	0.03	14.1	16.6	0.07
White (%)	17.7	21.2	0.09	17.5	20.5	0.08
Other (%)	6.1	3.9	0.10	6.4	3.9	0.11
# of Students	423	410		376	362	
Joint χ^2 p-value[2]		0.2411			0.2142	

[1] Columns under the subheading "Randomized" include all students that were randomized at the beginning of the semester. Columns under the subheading "Completed Final" include only those students that completed the final at the end of the semester.

[2] Joint χ^2 p-value is obtained from a joint test of the coefficients in a logistic regression that includes an indicator for group assignment as the dependent variable and the characteristics with group means imputed for missing characteristics and indicators for missing characteristics as the independent variables.

Table 3: First stage effects of nudging, grading, and the interaction on attempts, total time spent given attempting, and score given attempting[1]

	Attempt				**Total Time**		**Score**	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Graded	71.8***	71.2***			31.4***	30.6***	48.8***	49.3***
	(1.3)	(1.2)			(1.9)	(1.5)	(2.2)	(1.8)
Nudged	3.0***		3.0***		3.5**		-2.9	
	(0.8)		(0.8)		(1.8)		(2.6)	
Graded×Nudged	-1.1		-1.1		-0.6		0.6	
	(1.0)		(1.0)		(2.0)		(2.6)	
Worth 1%[2]			69.1***	68.6***				
			(1.5)	(1.4)				
Worth 2%			73.7***	73.2***				
			(1.4)	(1.3)				
Worth 3%			72.4***	71.9***				
			(1.4)	(1.3)				
Untreated Mean	12.6	12.6	12.6	12.6	10.8	10.8	38.7	38.7
# of Obs	9,996	9,996	9,996	9,996	4,857	4,857	4,959	4,959
# of students	833	833	833	833	787	787	787	787

[1] Coefficients are estimated via OLS. Standard errors in parenthesis are clustered by student. All models include student fixed effects. Assignment instead of chapter fixed effects are included in even numbered models. Significance levels are indicated by $* < .1$, $** < .05$, $*** < .01$.

[2] The worth of a assignment denotes how much students final grade the assignments were worth. All graded assignment effects are measured relative to recommending the same assignments.

Table 4: Reduced form and IV effects of grading and attempting assignments, on answering related exam questions correctly[1]

	Reduced Form[2]			**IV[3]**		
	(1) All	**(2)** Midterm	**(3)** Final	**(4)** All	**(5)** Midterm	**(6)** Final
Graded	2.6***	1.3	4.0***			
	(1.0)	(1.4)	(1.5)			
Attempted				3.4***	1.7	5.1***
				(1.2)	(1.7)	(1.8)
Ungraded Mean	52.8	60.4	44.9	52.8	60.4	44.9
# of Obs	9,072	4,644	4,428	9,072	4,644	4,428
# of Students	785	778	738	785	778	738

[1] All models have student and assignment fixed effects. Standard errors in parenthesis are clustered

[2] Coefficients in reduced form are estimated using OLS.

[3] All models have student and assignment fixed effects. Standard errors in parenthesis are clustered by student. Significance levels are indicated by $* < .1$, $** < .05$, $*** < .01$.

Table 5: First stage effects of nudging and grading on attempts by baseline characteristics[1]

	Race			Math SAT Score			Gender		GPA	
	Asian	White	Other	< Med[2]	> Med	Miss[3]	Female	Male	< Med	> Med
Graded	73.5^{***}	75.6^{***}	66.0^{***}	71.1^{***}	75.5^{***}	65.0^{***}	72.7^{***}	70.6^{***}	68.0^{***}	75.7^{***}
	(1.7)	(2.7)	(2.4)	(1.9)	(1.8)	(3.1)	(2.3)	(1.6)	(1.8)	(1.6)
Nudged	3.5^{***}	2.9	2.3	1.9^{*}	4.1^{***}	2.9	5.2^{***}	1.5	0.4	5.7^{***}
	(1.1)	(1.8)	(1.5)	(1.1)	(1.3)	(2.1)	(1.4)	(1.0)	(0.9)	(1.3)
Graded×Nudged	-1.8	-3.1	1.3	1.6	-3.1^{**}	-2.0	-3.2^{*}	0.4	2.0	-4.6^{***}
	(1.3)	(2.3)	(1.9)	(1.5)	(1.6)	(2.5)	(1.8)	(1.3)	(1.4)	(1.5)
Untreated Mean	13.5	10.0	12.6	10.7	12.2	17.5	16.1	10.2	8.2	16.9
# of Obs	5,160	1,944	2,892	4,068	4,044	1,884	2,832	5,580	4,896	4,848
# of students	430	162	241	339	337	157	236	465	408	404

[1] Coefficients are estimated using OLS. All specifications are clustered by student. All models have student and chapter fixed effects. Standard errors in parenthesis are clustered by student. Significance levels are indicated by $* < .1$, $** < .05$, $*** < .01$.

[2] Med refers to median.

[3] Miss refers to the SAT score being missing for that group.

Table 6: IV effects of attempting assignments on answering related exam questions correctly by baseline characteristics[1]

	Race			Math SAT Score			Gender		GPA	
	Asian	White	Other	< Med[2]	> Med	Miss[3]	Female	Male	< Med	> Med
Attempted	2.9*	7.2***	1.4	5.2**	4.7***	-4.2	5.2**	2.5	4.1**	2.9*
	(1.7)	(2.7)	(2.5)	(2.0)	(1.8)	(3.2)	(2.4)	(1.6)	(1.8)	(1.7)
Ungraded Mean	54.7	52.9	49.3	46.8	57.9	54.7	54.6	50.3	45.7	59.3
# of Obs	4,772	1,736	2,564	3,644	3,764	1,664	2,636	4,940	4,228	4,628
# of students	410	151	224	315	325	145	226	432	373	393

[1] Coefficients are estimated using 2SLS. All specifications are clustered by student. All models have student and chapter fixed effects. Standard errors in parenthesis are clustered by student. Significance levels are indicated by $* < .1$, $** < .05$, $*** < .01$.

[2] Med refers to median.

[3] Miss refers to the SAT score being missing for that group.

Table 7: First stage effects of nudging and grading on total time on all attempts given attempting by baseline characteristics[1]

	Race			Math SAT Score			Gender		GPA	
	Asian	White	Other	< Med[2]	> Med[3]	Miss	Female	Male	< Med	> Med
Graded	30.8***	27.7***	34.5***	32.3***	26.4***	39.0***	34.2***	28.2***	30.3***	31.3***
	(2.1)	(2.6)	(4.1)	(3.6)	(1.9)	(3.6)	(2.7)	(2.8)	(2.9)	(2.2)
Nudged	5.0**	2.4	0.7	-0.1	1.1	12.0***	6.7***	-1.3	-0.9	5.1**
	(2.2)	(3.0)	(3.4)	(3.0)	(2.3)	(3.0)	(2.6)	(2.6)	(2.7)	(2.1)
Graded×Nudged	-1.7	2.3	0.3	2.9	3.8	-13.5***	-2.7	3.5	1.7	-0.7
	(2.4)	(3.4)	(3.9)	(3.4)	(2.6)	(3.5)	(3.0)	(2.9)	(2.9)	(2.4)
Untreated Mean	9.7	13.2	11.7	12.0	10.7	9.4	10.2	11.1	8.7	11.9
# of Obs	2,612	911	1,334	1,883	2,035	939	1,512	2,524	2,051	2,687
# of students	410	151	226	322	322	143	229	430	370	398

Note: Coefficients are estimated via OLS. All specifications are clustered by student. All models have student level fixed effects. Standard errors in parenthesis are clustered by student. Significance levels are indicated by $* < .1$, $** < .05$, $*** < .01$. Miss refers to the SAT score being missing for that group, and Med refers to median.

[1] Coefficients are estimated using OLS. All specifications are clustered by student. All models have student and chapter fixed effects. Standard errors in parenthesis are clustered by student. Significance levels are indicated by $* < .1$, $** < .05$, $*** < .01$.

[2] Med refers to median.

[3] Miss refers to the SAT score being missing for that group.

Table 8: First stage effects of nudging and grading on score given attempting by baseline characteristics[1]

	Race			Math SAT Score			Gender		GPA	
	Asian	White	Other	$<$ Med[2]	$>$ Med	Miss[3]	Female	Male	$<$ Med	$>$ Med
Graded	48.2***	53.1***	47.8***	52.2***	47.6***	45.4***	41.6***	53.8***	53.9***	46.8***
	(2.8)	(5.3)	(3.8)	(3.4)	(3.6)	(3.5)	(3.5)	(2.8)	(3.3)	(2.6)
Nudged	-4.4	8.9*	-6.2	-3.7	-7.1*	5.1	-9.4**	0.0	-3.9	-3.1
	(3.2)	(5.3)	(4.3)	(3.8)	(3.8)	(4.5)	(3.9)	(3.6)	(3.6)	(3.0)
Graded×Nudged	2.4	-11.0**	3.4	0.3	5.8	-7.1	7.2*	-2.0	0.9	1.2
	(3.2)	(5.4)	(4.4)	(3.9)	(3.8)	(4.7)	(3.9)	(3.7)	(3.8)	(3.0)
Untreated Mean	39.9	37.7	36.9	32.3	43.8	39.3	47.1	30.8	28.2	43.3
# of Obs	2,658	931	1,370	1,933	2,067	959	1,535	2,584	2,093	2,744
# of students	410	151	226	322	322	143	229	430	370	398

[1] Coefficients are estimated using OLS. All specifications are clustered by student. All models have student and chapter fixed effects. Standard errors in parenthesis are clustered by student. Significance levels are indicated by $* < .1$, $** < .05$, $*** < .01$.

[2] Med refers to median.

[3] Miss refers to the SAT score being missing for that group.

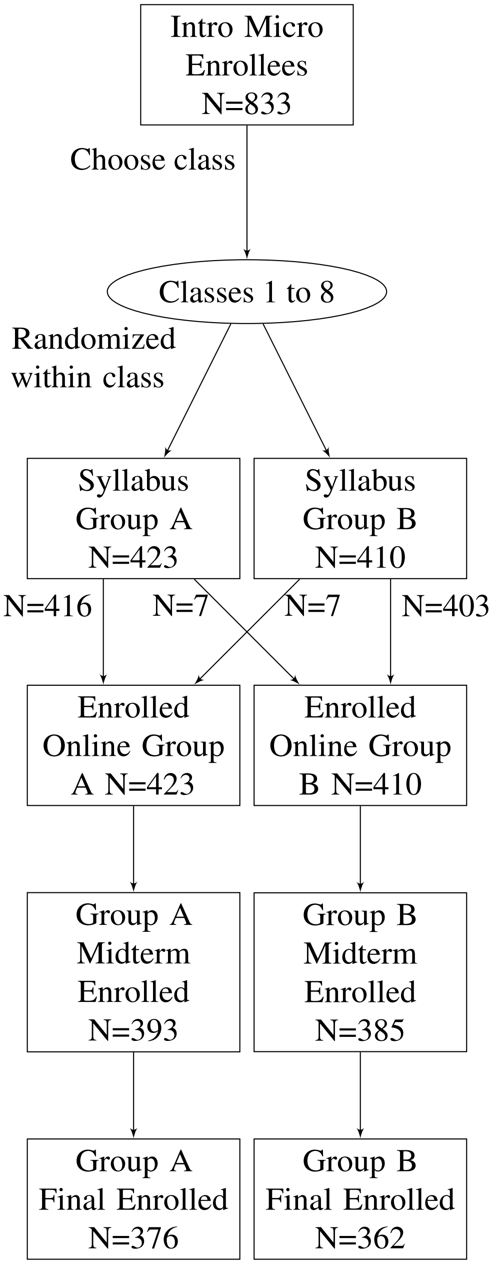

Intro Micro Enrollees N=833
Choose class
Classes 1 to 8
Randomized within class
Syllabus Group A N=423
Syllabus Group B N=410
N=416
N=7
N=7
N=403
Enrolled Online Group A N=423
Enrolled Online Group B N=410
Group A Midterm Enrolled N=393
Group B Midterm Enrolled N=385
Group A Final Enrolled N=376
Group B Final Enrolled N=362

Figure 1: Experimental Flowchart

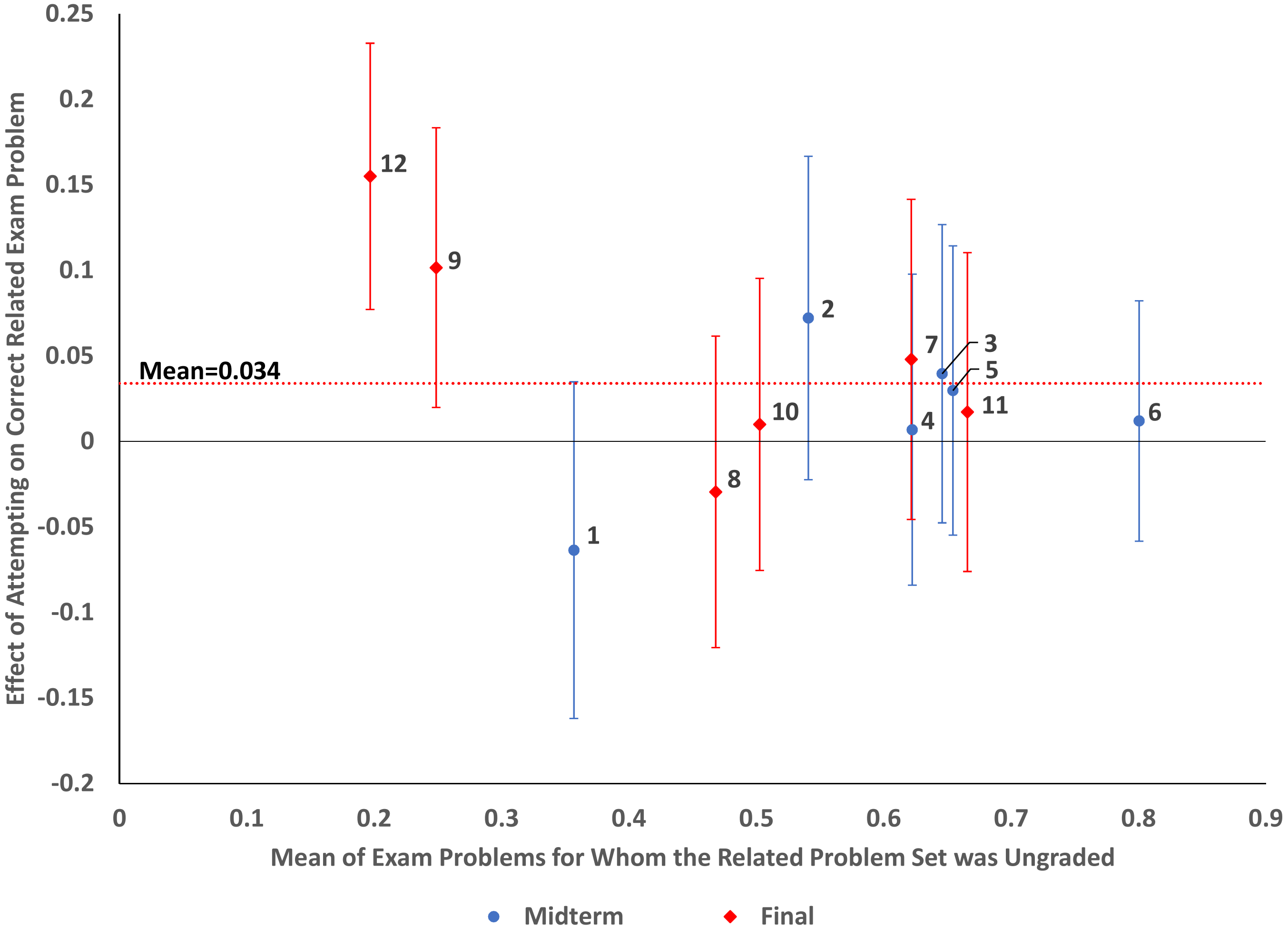


Figure 2: Problem Specific effects and confidence intervals by mean on exam

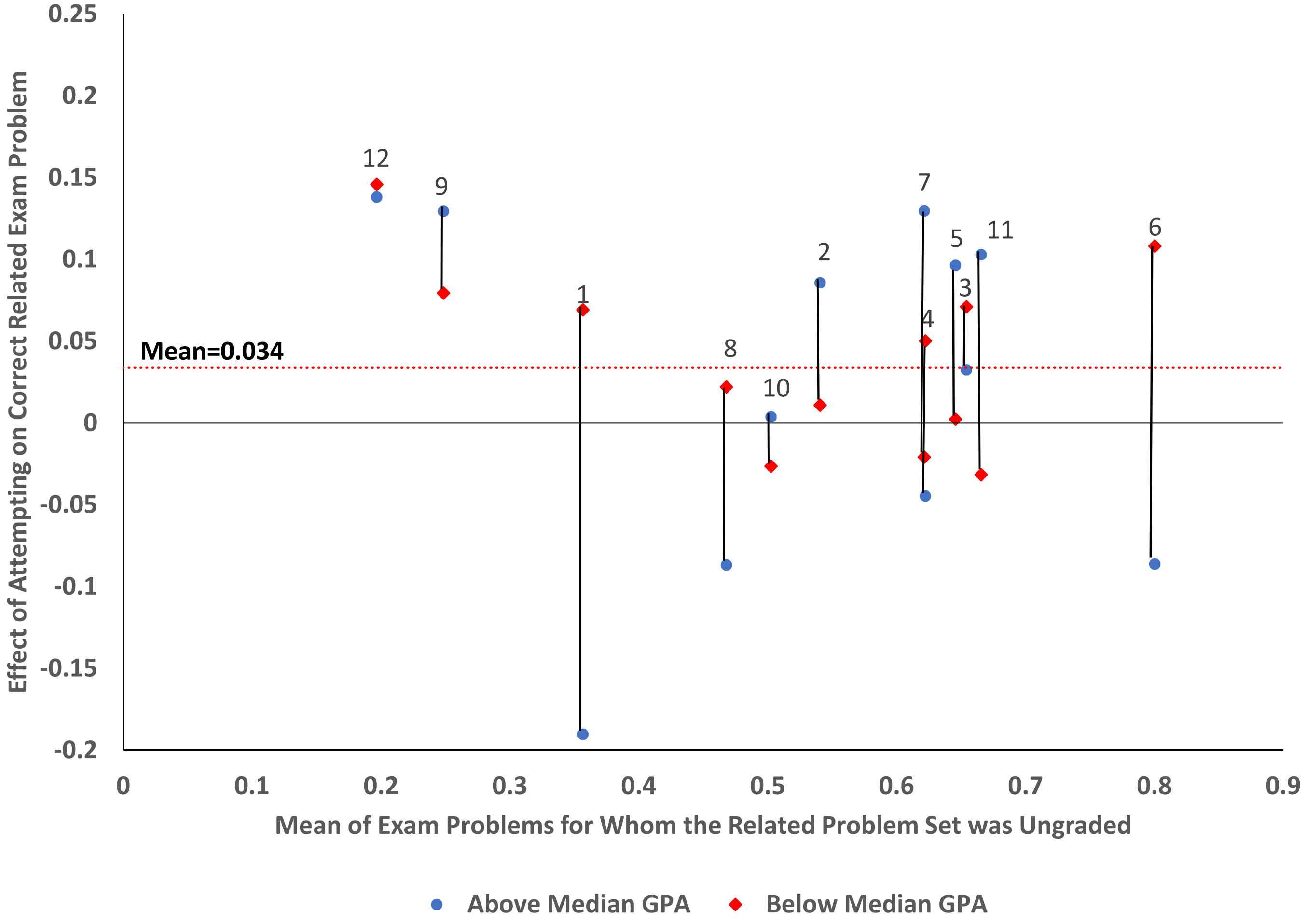


Figure 3: Problem specific effects by ungraded mean on exam and GPA

Appendix A: A model of student effort

A.1 Objective Function

I start from the assumption that students are solely interested in obtaining a grade[11] (S) which yields a benefit at the rate of w, which captures all monetary and non-monetary compensation for the higher grade. I will also assume that there are three things that determine S: ability (A), classroom capital (K), and allocated time to homework in the class (E). I will also assume that students choose E, which is costly and possibly related to A, to maximize the following objective function

$$wS(E, K, A) - c(E, A). \tag{4}$$

I will assume A and K are held fixed upon selection of a classroom and abstract away from them. S is defined as an identity by two components, performance on exams (X) which are a function of E, and performance on homework (Q)[12] is a function of E for a give student so that

$$S = (1 - \Lambda)X(E) + \Lambda Q(E). \tag{5}$$

Where Λ equals the share of a student's grade that belongs to all homework and is a number between 0 and 1. I can break Q up into independent problem sets so that

$$Q = \sum_{t=1}^{n} \lambda_t q_t(e_t). \tag{6}$$

In this case, t indexes the problem set, and n is the number of problem sets. λ_t equals the share that each problem set is worth towards Q. The grade on each problem set is of course determined by e_t. X is itself a function of e_t on each of these problem sets with a weight of ρ_t on the exam so that

[11] I choose to use grades instead of human capital (H) as was used in a similar model in Dee and Jacob (2012). The value of grading could be the human capital formation it represents or the signal it sends to the market. I remain agnostic as to which is more important as it does not affect any of the predictions from my model.

[12] I use homework here for simplification purposes, but any learning activities which contribute to retention of knowledge students use on the exam could apply here as well, including attendance, participation, and readings.

$$X = \sum_{t=1}^{n} \rho_t x_t(e_t). \tag{7}$$

ρ_t represents the fraction of worth on all exams related to problem set q_t and multiplies a function of e_t. Students lack information about whether the size of ρ_t is zero or some positive number. They may ask the question "Professor, will this be on the exam?" to obtain more information.

I will also assume that the cost functions for each problem set is separable so that

$$C = \sum_{t=1}^{n} c_t(e_t). \tag{8}$$

I can plug the equations for C, Q and X back into the objective function. I get

$$w[(1-\Lambda)\sum_{t}^{n} \rho_t x_t(e_t) + \Lambda \sum_{t}^{n} \lambda_t q_t(e_t)] - \sum_{t}^{n} c_t(e_t). \tag{9}$$

A.2 Optimization and predictions

Note that x_t, q_t and c_t are all strictly increasing functions with respect to e_t. Now I differentiate with respect to the choice variable e_t for t=j and set equal to 0. Note that because all functions with e_t are linearly separable, all terms where $t \neq j$ drop out of the model and I am left with

$$w[(1-\Lambda)\rho_j \frac{\partial x_j}{\partial e_j} + \Lambda\lambda_j \frac{\partial q_j}{\partial e_j}] = \frac{\partial c_j}{\partial e_j}. \tag{10}$$

Since all function are increasing in e_j, assuming x_j and q_j are concave functions in e_j and that c_j is a convex function in e_j[13], this equation has at least one positive solution in e_j and this solution is increasing in ρ_j and λ_j. These are the parameters I manipulate for this experiment. For example, by telling students a problem is likely to be on the exam, I increase ρ_j. Further, when a problem is graded it increases λ_j from zero to $1/25^{th}$, $2/25^{ths}$ or $3/25^{ths}$.[14]

[13] These are standard assumptions in a cost benefit framework.

[14] Of course, student expectations about ρ_t might also be a function of λ_t since students may take it as additional information of how important professors perceive a problem set to be for X. The purpose of this experiment is in giving explicit information about the probability of including a particular type of problem on a test, altering only ρ_t versus altering λ_t even if altering λ_t has a tertiary effect of also altering ρ_t. Altering λ_t is a proven method for increasing work

A.3 Ordering expectations

I will order my expected responses in effort with respect to grading and nudging interventions. In this experiment, each question on an exam is worth approximately 1.125 percentage points towards a student's final grade, which I will round down to 1 percentage point for simplification purposes. Let's suppose that ρ_j moves from zero to one when students are told problems like it are likely to appear on the exam. I make the homework problem sets worth a minimum of 1 percentage point of their final grade, so $(1 - \Lambda)\rho_j$ and $\Lambda\lambda_j$ are equal when a homework problem set is worth 1 percentage point.

Therefore the relative payoff to students in the nudging and grading interventions are approximately $w[\frac{\partial x_j}{\partial e_j}]$ and $w[\frac{\partial q_j}{\partial e_j}]$ when homework problems are worth 1 percentage point towards the final grade. The relative behavioral changes are therefore approximately related to the relative sizes of $\frac{\partial x_j}{\partial e_j}$ and $\frac{\partial q_j}{\partial e_j}$. A priori, there is good reason to believe that $\frac{\partial q_j}{\partial e_j} > \frac{\partial x_j}{\partial e_j}$ since students are allowed unrestricted time, access to their textbooks, the internet, and whatever else they need while completing problem sets. By contrast, students must work on their exams using only the knowledge they have brought with them from prior work in a limited time frame.

Student beliefs about $\frac{\partial q_j}{\partial e_j}$ and $\frac{\partial x_j}{\partial e_j}$ may also influence how students respond to increases in λ_j and ρ_j. For example, for students that have experienced some success with how their time input influences their grade in the past, they may be more responsive to information about whether a problem will be on the test and whether a problem set is graded or not. In addition, students may have much different function of $\frac{\partial c_j}{\partial e_j}$, which may influence how they respond to interventions to increase time input.

Appendix B: Waiver of consent

During the IRB process, I sought for and received a waiver of consent. The following are the conditions any study needs to meet in order to obtain this waiver. It must involve no more than

while altering only ρ_t is less studied. In addition, altering λ_t in the same experiment allows for recovery of effects in exam performance since large first stage effects are expected.

minimal risk to subjects. The research could not be carried out practicably without the waiver. The waiver or alteration will not adversely affect the rights and welfare of the subjects. Finally, the subjects will be provided with additional information about their participation if the research involves deceptive practices.

This research involved no more than minimal risk to subjects since they would have been required to complete the same amount of work for a grade and received the same number of recommended problems if the research was not conducted. Care was taken to ensure that students received the same level of difficulty in problems on average.[15]. In addition the treatment effects of assignment to A or B on answering the set of 12 questions correctly was very close to zero and insignificant.

It could not be carried out practicably without the waiver because informing subjects about their participation in the study would have put the stable unit treatment value assumption at greater risk of being violated. The reason is that subjects might have become aware of their assignment by talking to one another, and so have been more likely to complete problems that were graded for the opposite group.

The rights and welfare of subjects were not impeded because both groups received the same treatment on different problems with similar levels of difficulty. Their selection into either group was therefore equitable since both groups selection was equitable. My assessments of risk at the midterm and final revealed no differential impacts on group A and group B on their test scores, nor at the end of the semester on their final grades. In Table B1 you can see all assessed outcome variables and that they are not significant by assignment group. Further, the privacy and confidentiality of subjects was maintained by creating a test score key which I could link to their online work output and characteristics using a secure encrypted folder on a computer not connected to the internet and than discard identifiable information for further analysis.

Finally, the research was in no way deceptive because their syllabus contained the following

[15] Indeed the ungraded exam means for sets of problems by groups were not significantly or practicably different from one another, with 53.2% correct corresponding problems for group A and 52.2% correct corresponding problems for group B

Table B1: Effect of assignment to group B on attempting, time on first attempt, total time spent, score given attempting, and correct answer on exam[1]

	Attempt	# of Attempts	Time on 1st	Total Time	Score	Correct on Exam
Group B	1.6	-0.1	-0.8	-1.6	2.2	1.0
	(1.4)	(0.1)	(1.4)	(1.6)	(2.4)	(1.3)
Untreated Mean	12.6	1.4	8.7	10.8	38.7	60.6
# of Obs	9,996	4,664	4,857	4,857	4,959	9,846
# of Students	833	833	787	787	787	785

[1] Coefficients are estimated using OLS. Standard errors in parenthesis are clustered by student. All models have student fixed effects. Significance levels are indicated by $* < .1$, $** < .05$, $*** < .01$.

statement in italics. "Please note that in some post-lecture quizzes we give different questions of equal difficulty to different students. We vary the questions so that students see a range of problems." This informed the students that they would receive different questions without revealing that they were involved in an experiment.

Appendix C: Treatments by group and problem set

Table C1 shows the intended treatment. There were two exceptions to this plan. For experimental chapter 5, there was a snow day and the professor changed their schedule without informing the course administrator before doing so. The scheduled assignment for this chapter occurred earlier then was intended and no nudge was sent for this chapter as a result.

Second due to a simple coding error on my part, another class had its nudges flipped in the last two chapters of the semester, with group A getting nudged for their recommended problem and group B getting nudged for their graded problem. In chapter 6, group A was nudged for their recommended problem set and group B was nudged for their graded problem set. It should be noted however, that neither of these mishaps led to any detectable harm, as the effort response from nudges was small, and the difference between graded and recommended problem nudges even smaller. Re-running the results without these classes included does not alter the results in any table or figure in any substantial way.

Table C1: Interventions by problem set and group

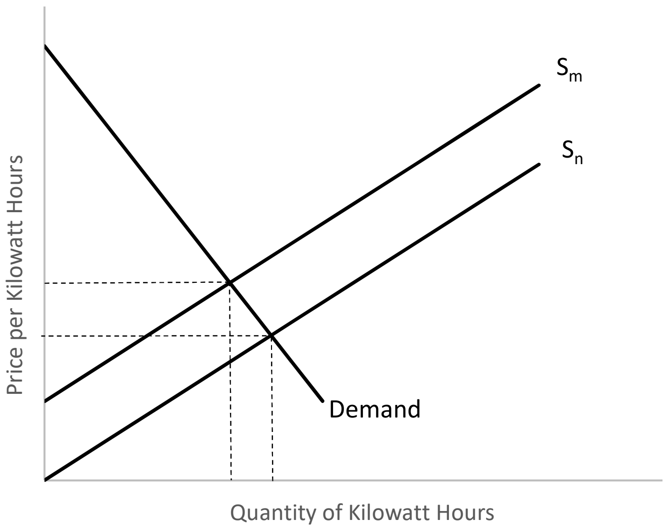
Chapter/Problem Set	Group A Treatment	Group B Treatment
Chapter 1 Problem Set 1	Graded(1%)/Nudged[1]	Recommended/Nudged
Chapter 1 Problem Set 2	Recommended/Not Nudged	Graded(1%)[2]/Not Nudged
Chapter 2 Problem Set 3	Graded(2%)/Not Nudged	Recommended/Not Nudged
Chapter 2 Problem Set 4	Recommended/ Nudged	Graded(2%)/Nudged
Chapter 3 Problem Set 5	Graded(3%)/Nudged	Recommended/Nudged
Chapter 3 Problem Set 6	Recommended/Not Nudged	Graded(3%)/Not Nudged
Chapter 4 Problem Set 7	Graded(1%)/Not Nudged	Recommended/Not Nudged
Chapter 4 Problem Set 8	Recommended/ Nudged	Graded(1%)/Nudged
Chapter 5 Problem Set 9	Graded(2%)/Nudged	Recommended/Nudged
Chapter 5 Problem Set 10	Recommended/Not Nudged	Graded(2%)/Not Nudged
Chapter 6 Problem Set 11	Graded(3%)/Not Nudged	Recommended/Not Nudged
Chapter 6 Problem Set 12	Recommended/ Nudged	Graded(3%)/Nudged

[1] Nudged refers to telling a student a problem like their graded or recommended problem set is likely to be on the test.

[2] The 1%, 2%, and 3% indicate the percent towards final grade a graded problem is worth.

Appendix D: Test questions related to experiment

Figure: Kilowatt Hours



1. Pictured in the figure above are the supply and demand curves for in home energy. Due to activist worries, the government decides to ban nuclear power as a means of energy production, causing the supply curve to shift along an inelastic demand curve. Which way does the supply curve shift? What effect does this change have on total energy expenditure?

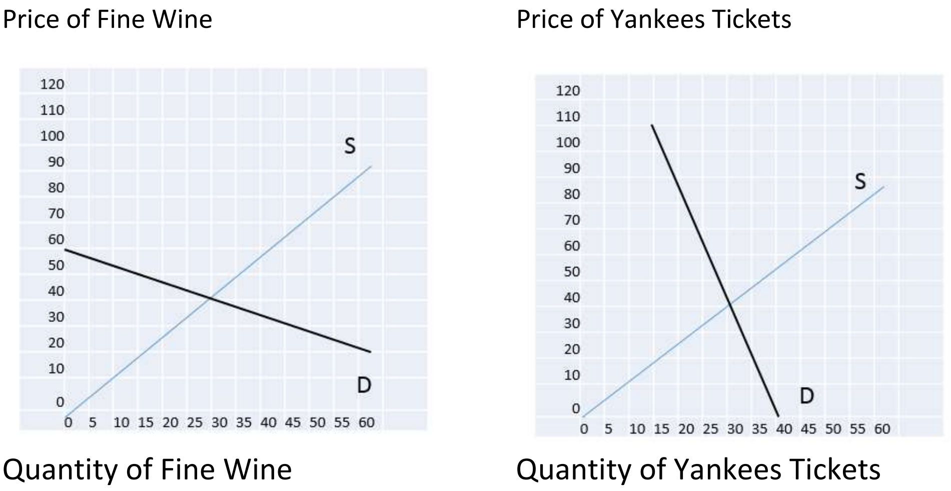
a. Right, increases

b. Right, decreases

3. Assume the following demand, $Q_d = 240 - 4P$ and supply equation, $Q_s = 4P$. Suppose the government taxes consumers T dollars such that the new demand equation is $Q_d = 240 - 4(P + T)$. What are the new equilibrium price and quantities?

a. P= 30-T and Q = 120-2T
b. P= 30-1/2T and Q = 120-2T
c. P= 120+T and Q= 30 +T
d. P= 120-T and Q = 30-T

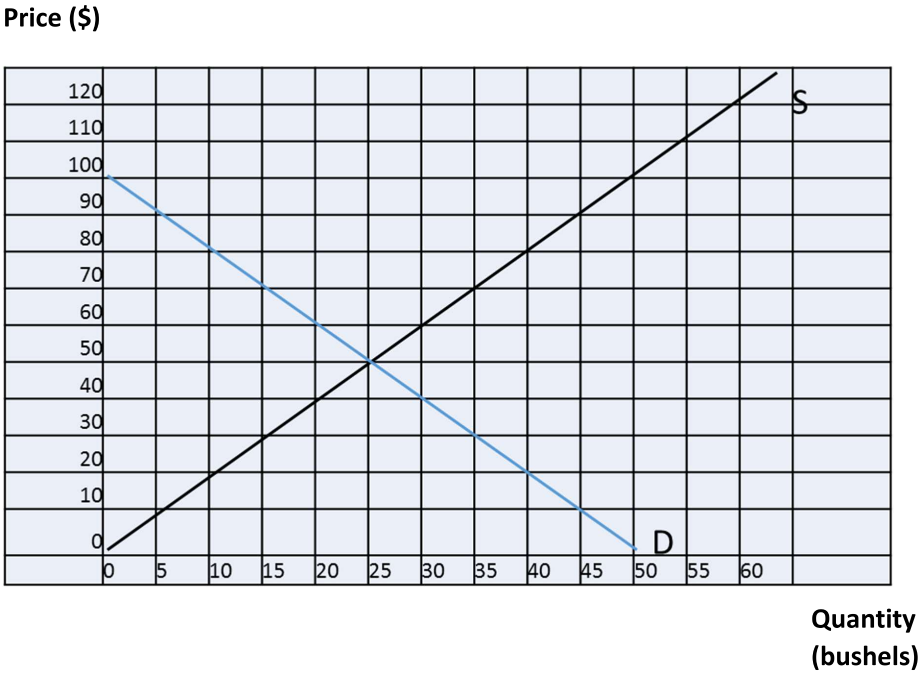
Figure: Fine Wine & Yankees Tickets



4. Refer to the figure above. Suppose the government imposes a $30 tax on both fine wine and Yankees tickets. Tax revenue (TR) will be ______ and dead weight loss (DWL) _______.

a. lower, lower in the market for fine wine
b. higher, higher in the market for fine wine
c. higher, lower in the market for Yankee tickets
d. lower, higher in the market for Yankee tickets

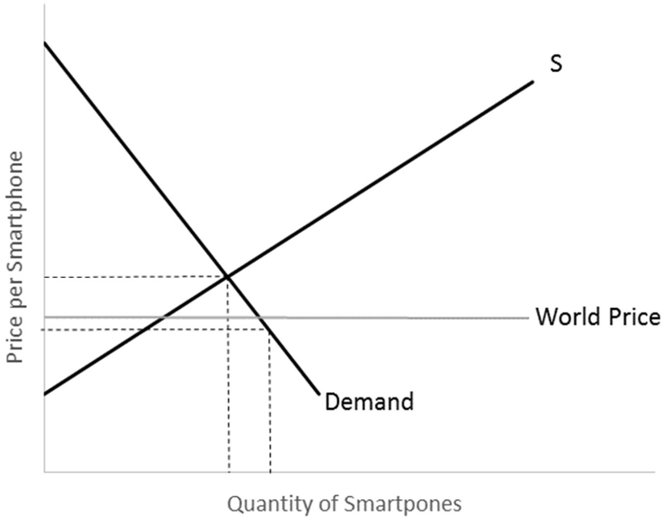
Figure: Avocados



5. The figure above shows the domestic supply and demand for avocados in the US. Suppose that the world price for avocados is $30 a bushel. The Trump administration wants to impose an import tariff of $10 per bushel. As a result of the tariff, the change in consumer surplus (CS) and producer surplus (PS) are…

a. CS falls by $200; PS rises by $200
b. CS falls by $325; PS rises by $175
c. CS falls by $100; PS also falls by $100
d. CS falls by $250; PS rises by $400

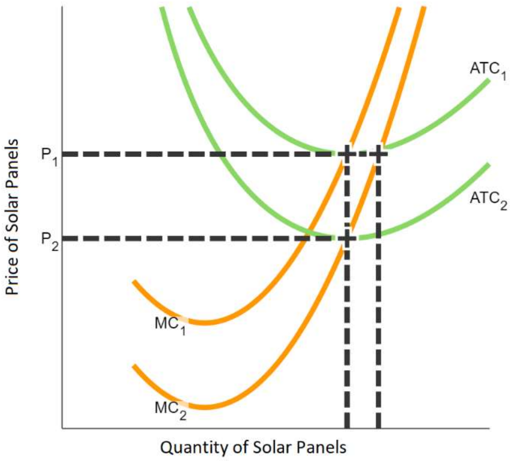
Figure: Smartphones



6. The figure above shows the supply and demand curves for smartphones produced and consumed in the US, as well as the world price for smartphones. A labor strike at a foreign production facility slows the production of smartphones overseas. What happens to the world price for smartphones? Who will be made better off?

a. Rise; US producers of smartphones
b. Rise; US consumers of smartphones
c. Fall; Foreign Producers
d. Fall; US consumers of smartphones

Figure 2



7. Refer to figure 2. In the year 2035 TechX discovers a way to mine rare earth metals from asteroids in close proximity to earth at virtually no cost. These rare earth metals are used in manufacturing solar panels. TechX is the only company with rockets this advanced and no company will be able to duplicate this for a long time. The following graph shows the marginal cost curve (MC1) and Average Total Cost Curve (ATC1) for everyone else in the short run and the marginal cost curve (MC2) and Average Total Cost Curve (ATC2) for TechX to produce Solar Panels. What happens to profits/losses for solar panel sale in the short run and the long run?
   a. TechX takes Economic Losses; TechX makes Economic Profit
   b. TechX makes Economic Profit; TechX takes Economic Losses
   c. TechX makes Economic Profit; TechX makes zero profit
   d. TechX makes zero profit; TechX makes Economic Profit

8. Suppose that each firm in a competitive industry has the following cost curves:
   Total cost: TC = 32 + ½ Q^2; where Q is the individual firm's quantity produced. MC=Q. Assume the market price is $14 per unit. If the market price falls, how much will each firm produce in the long run?
   a. 32
   b. 8
   c. 11
   d. 64

Scenario 1

A company is considering building a bridge across a river. The company would have monopoly control over the revenue and profit. The bridge would cost $1 million to build and nothing to maintain. The following table shows the company's anticipated demand over the lifetime of the bridge:

Price ***(Dollars per crossing)***	**Quantity** ***(Thousands of crossings)***
8	0
7	50
6	100
5	150
4	200
3	250
2	300
1	350
0	400

9. **Refer to Scenario 1.** If the company declined to build the bridge, should the government build it?
   a. Yes because the efficient number of crossings is 200
   b. No, because like the company, it would lose money
   c. Yes, because total surplus to society exceeds the costs
   d. No, because even where price equals marginal cost, the government would lose money

Scenario 2: Consider a monopolist with the following cost and demand curve. Concerned about high prices the government breaks up the monopolist and makes the industry competitive.

Demand:	$P=19-Q$
Total Cost:	$TC=1+Q+0.5Q^2$
Marginal Cost:	$MC=1+Q$

10. **Refer to Scenario 2:** What is the deadweight loss associated with the monopolist?
    a. $36
    b. $45
    c. $9
    d. $27

Scenario 6: Pete's is a small coffee company that is considering entering a market dominated by Starbucks. Each company's profit depends on whether Pete's enters and whether Starbucks sets a high price or a low price:

		Starbucks	
		High Price	Low Price
Pete's	Enter	**$0.5 million**, **$3 million**	**$2 million**, **$2 million**
	Don't Enter	**$1 million**, **$4 million**	**$0**, **$2.5 million**

11. **Refer to Scenario 6:** Which of the following best describes the likely equilibrium if any.
    a. Starbucks charges a low price and Pete's enters the market
    b. A dominant strategy that results in a Nash equilibrium is for Starbucks to charge a high price and for Pete's to enter the market
    c. The Nash equilibrium is for Pete's not to enter and for Starbucks to charge a high price
    d. There is no Nash equilibrium

Scenario 7: Consider a town in which only two companies, Agua and Eau, own wells that produce bottled water. Agua and Eau can pump and sell as much water as they want at no cost. For them, total revenue equals profit. The following table shows the town's demand schedule for water.

Price	**Quantity Demanded**	**Total Revenue**
(Dollars per gallon)	***(Gallons of water)***	***(Dollars)***
10	0	0
9	30	$270.00
8	60	$480.00
7	90	$630.00
6	120	$720.00
5	150	$750.00
4	180	$720.00
3	210	$630.00
2	240	$480.00
1	370	$370.00

12. **Refer to Scenario 7:** Agua and Eau have colluded for years to maximize profits. Agua's new ownership decides to break that arrangement and produce more bottled water. How low will the price fall as the two firms compete on output?
    a. $1
    b. $2
    c. $3
    d. $4

Table D1: Differences between experimental assignments and related test questions

Chapter/Assignment	Change	Effect Size
Chapter 1 Assignment 1	Switched inelastic supply curve with elastic supply curve	-0.06
Chapter 1 Assignment 2	Made quantity rise instead of stay same and price stay same instead of falling	0.07
Chapter 2 Assignment 3	Same Setup	0.04
Chapter 2 Assignment 4	Same Setup	0.01
Chapter 3 Assignment 5	Asked to compute exact change of consumer surplus and producer surplus	0.03
Chapter 3 Assignment 6	Adverse rather than positive supply shifting event	0.01
Chapter 4 Assignment 7	Same setup	0.05
Chapter 4 Assignment 8	Skips lead up steps	-0.03
Chapter 5 Assignment 9	Same Setup	0.10
Chapter 5 Assignment 10	Asks about deadweight loss	0.01
Chapter 6 Assignment 11	Same Setup	0.02
Chapter 6 Assignment 12	Asks for price instead of optimal output	0.16