COLLEGE TEACHING
https://doi.org/10.1080/87567555.2020.1865865

Routledge
Taylor & Francis Group

The Effect of Grades on Student Performance: Evidence from a Quasi-Experiment

Thomas Gray and Jonas Bunte

The University of Texas at Dallas

ABSTRACT

Grades are frequently understood as an output of the educational process. However, they may also be an input: receiving a midterm grade may affect student performance in the remainder of the semester. Utilizing a regression discontinuity design to isolate the causal impact of midterm grades on subsequent student performance in the same course, we find that low grades appear to incentivize students to improve their subsequent course performance. When analyzing how students accomplish these improvements, our results indicate that students do not focus on studying for final exams, presumably because this is a risky strategy of "saving" the final course grade after a low midterm grade. Instead, we find that students improve their final course grade by increasing their performance on low-stakes assessments, such as participation, reading quizzes, and in-class clicker exercises. Additional analysis reveals that this effect of midterm grades is particularly strong among men, and to a lesser degree among younger students as well as nonsocial science majors. Our findings have practical implications for instructors for how they can motivate students in the second half of the semester.

KEYWORDS

Grades; regression discontinuity; student performance

Grades are an unavoidable part of most of collegiate educating. Though they are frequently understood as an output of the educational process, they may also be an input: receiving a low grade in the middle of a semester may affect a student's subsequent effort and learning in the remainder of the term. Theoretically, receiving a low grade in the middle of the semester might have positive or negative effects on a students' subsequent course performance. On the one hand, a low grade might signal students that they need to study harder and devote more time to this course, resulting in improved performance in the second half of the course. On the other hand, it is also possible that a low grade may discourage students, causing them to disengage from the course which presumable results in lower grades.

We analyze how Midterm Grades (hereinafter "Mid-Term Grades [MTGs]") given based on the students' work in the first half of the semester affect students' subsequent performance in that course. MTGs are given approximately halfway through the course based on all work completed up until that point, based on the grading system that is used for the eventual overall course grade. We examine the effect of MTGs in the context of an introductory survey of American national politics taught at a public research university in the American southwest in Fall 2017.

We find that students who receive lower MTGs appear to improve their performance in the remainder of the course, a pattern not found among those who received the higher MTGs. We find that this is pattern is particularly strong among students who are close to the cutoff for high grades (between the letter-grades A and B) as well as passing (the letter-grades C and D), while the difference is less pronounced between middling grades (the letters B and C).

We further investigate several potential mechanisms through which MTGs might shape students' behavior. A low MTG might lead students to spend more time on the remaining exams; conversely, students may focus more on the small, low-stakes assessments as well as participation; or both. We find that low MTGs do not affect final exam performance but significantly increase performance on low-stakes assignments and participation. Lastly, we examine how the identified effects are conditioned by student characteristics. For example, we find that this effect of midterm grades is

CONTACT Thomas Gray thomas.gray1@utdallas.edu University of Texas at Dallas, Political and Policy Sciences, 800 West Campbell Road, GR 31, Richardson, Texas, 75080-3021, USA.

particularly strong among men, but not among women.

This article makes several contributions. Substantively, we are the first to explicitly test how MTGs may shape behavior of students, rather than merely identifying the effect of MTGs on final course grades. Further, our analysis is unique in examining how such behavior changes are conditional on student characteristics. Methodologically, our research design can address issues of selection bias in ways previous work was unable to do. Further, we utilize causal inference methodologies to address potential bias due to unobservable student characteristics such as their innate abilities. Our research has several practical implications for instructors, such providing support for mid-semester reflection papers and the way in which students should receive feedback on assignments.

The remainder of the article is structured as follows. The next section discusses our theoretical expectations for the effect of MTGs on the final course grade as well as student behavior. Section 2 describes two methodological issues—selection bias and unobservable student characteristics—that make identifying the causal effect of MTGs on subsequent student performance challenging, and discuss how our research design is able to address these challenges. We then present three sets of findings in Section 3: the effect of MTGs on the subsequent course performance, on the behavior of students, and how these effects depend on student characteristics. Section 4 discusses practical implications of our research for instructors; Section 5 concludes.

1. Theoretical expectations for the effect of grades on student performance

1.1. The effect of MTGs on subsequent student performance

A priori is not clear how midterm grades might shape performance over the remainder of the semester. One possibility suggests that receiving a low grade provides students with the motivation to increase their subsequent performance. Haladyna (1999) argues that most students view high grades as positive recognition of their success, and work hard to avoid the consequences of low grades. Subsequent empirical efforts confirm this view: Cameron and Pierce (1996) find that receiving a poor grade might prompt students to increase their efforts. Main and Ost (2014) show that the first exam letter grade can affect student performance on the second exam. Students who earn just below a B on the first exam outperform similar students on the second exam by about half a letter grade. Oettinger (2002) finds that students who are on grade boundaries prior to the final exam are more likely to score higher on the final exam compared to other students.

We build upon this expectation by pointing to an implicit assumption that has not yet received sufficient attention: if low grades increase student performance, the opposite might be true as well. That is, receiving a high grade might lead to a decline of effort as students might think they have accomplished mastery of the material already. Both effects might be operating simultaneously to result in a significant difference in subsequent student performance. Our first hypothesis suggests

> **Hypothesis 1**: In comparing students, those with low mid-semester grades should exhibit better subsequent grades than students with high grades, while the opposite is the case for students with high grades.

A second possibility, however, suggests the opposite relationship: Low grades might discourage students from exerting more effort, thereby leading to a deterioration of performance in the remainder of the semester. Conversely, high grades might reinforce students' study behavior, resulting in more devotion to their studies.

Consistent with this argument, some work suggests that only positive feedback enhances motivation, while negative feedback does not (Deci, Koestner, and Ryan 1999; Dev 1997; Cameron and Pierce 1994). For example, negative feedback may lead undergraduates to spend less time on a task (Shanab et al. 1981, Hattie and Timperley 2007). Specifically, Guskey (2017) argues that low grades may discourage students and cause them to withdraw from learning. Selby and Murphy (1992) argue that students detach themselves from the course upon receiving a low grade in order to protect their self-images. A second possible hypothesis is thus

> **Hypothesis 2**: In comparing students, those with low mid-semester grades should exhibit worse subsequent grades than students treated with high grades, while the opposite is the case for students with high grades.

1.2. The effect of MTGs on students' behavior

If grades lead to improved student performance, it would be important to know *how* students accomplish this. How does the signal of a low/high grade alter student behavior with respect to this class? Existing studies have focused primarily on identifying the

effect of grades on performance without testing the way in which students adjust their behavior that result in improved performance. For example, Oettinger (2002) as well as Main and Ost (2014) assume that improved grades are a reflection of increased student effort. But how does this increased student effort manifest itself?

We argue that grades may incentivize students to re-allocate their time and energy to particular tasks within the course. Given the signal of a low grade, students may decide to focus their efforts on one aspect of the course to maximize their final course grade. We see two possible ways in which grades may shape students' allocation of time within a course.

First, students may focus on studying for the final exam. If a student received a low MTG, the perceived importance of the final exam might increase: Students might over-estimate the potential of that final to "save" the final course grade, particularly if the final exam is given large weight in the final course grade calculations. The multiple low-stakes assessments may not be worth much attention anymore, given the small number of points that can be earned for each of them. Even though the cumulative number of points obtainable with these low-stakes assignments might still be significant, the "busy work" might not be perceived as a worthwhile investment of time.

Hypothesis 3: Receiving a high/low grade should lead to students exhibiting improved performance on the final exam.

Second, it is equally plausible that students focus on frequent low-stakes assessments. Such assignments include simple quizzes and reading checks, in-class attention checks, and basic participation in class. Two considerations may guide students' behavior. First, the degree of intellectual lifting required to successfully complete low-stakes assignments is significantly lower than that of a final exam. Students may interpret their low MTG as evidence that they are particularly well suited to perform successfully on exams in that course. In contrast, low-stakes assessments merely require commitment to complete these tasks, rather than intellectual efforts. Second, focusing on the final exam to "save" the final course grade is a high-risk, high-reward strategy as the possibility of "bombing" the final exam does exist. In contrast, focusing on multiple, smaller, low-stakes assessments might be perceived as preferable as it promises points with a higher degree of certainty.

Hypothesis 4: Receiving a high/low grade should lead to students exhibiting improved performance on the low-stakes assessments.

2. The challenges of identifying the causal effect of grades

The previous section developed four hypotheses linking MTGs to student performance. Empirically testing these hypotheses, however, is challenging. Two issues threaten the accuracy of inferences from statistical analyses, thereby undermining our ability to confidently determine which of the hypotheses are supported by the data. This section briefly describes the two challenges to identifying the causal effect of MTGs on student performance and discusses how our research design addresses both issues.

2.1. Selection bias

Selection bias may occur when the observations analyzed are not randomly drawn from a population, but instead systematically include too many individuals with certain characteristics while individuals with other characteristics are under-represented. This is a serious threat in the context of colleges as students typically self-select into a particular course.

For example, students might pick the classes where they expect to get the highest grades. Existing work provides evidence that students care about the grades they receive and they respond by taking classes where they will get easier grades (Bar, Kadiyali, and Zussman 2009) and switching out of majors that do not grade easily (Sabot and Wakeman-Linn 1991). This might imply that students who care about receiving high grades are over-represented. Due to the over-representation of students with increased sensitivity to grades, results from regression analysis will likely be biased upward and overstate the importance of MTGs for subsequent performance.

Previous studies examined the effect of grades on subsequent outcomes only among students that enrolled in a particular class by their own choice. For example, Oettinger (2002) as well as Main and Ost (2014) examine the effect of grades in economics classes. As students self-select into these classes, the resulting sample may not be representative. As a result, selection bias is likely.

To avoid selection bias, we ideally need a sample of students that is not affected by self-selection but instead closely resembles a random draw from the university's total student population. Our research design accomplishes this by examining data from a class that is mandatory for all students at this university: Every student, irrespective of their major, is required by the Texas legislature to take the class

"American National Government." As a result, our sample does not include students that self-selected into this course, but a truly representative sample of the average student population at our university. An added benefit is that we can examine whether the effect of MTGs on student performance differs across students enrolled in various majors, as all majors are represented in our sample. Our research design is capable of analyzing the effect of MTGs on student performance while accounting for possible selection bias.

The resulting sample that forms the basis of our analysis is representative of the student population at the university in general. Consistent with a STEM-dominated university, our sample is slightly male-dominated as 65% of our students were male. Similarly, all majors are represented in the sample. However, with 55% were majoring in a "STEM" field. Only three percent were political science majors. This accurately reflects the relative frequency of majors across the entire student population at the university.

One exception, however, is the age distribution. As a required course, the vast majority of students take this course in their first or second year at the university. About 65% were first-year students (compared to about 18%, 11%, and 6% for second-, third-, and fourth-year students). This means that our population is younger than the average college student at the university. Thus, our results should be understood as most applicable to new college students in large, introductory courses. However, we argue that possible selection bias based on students' age are likely substantially smaller than selection effects due to self-selection into particular courses.

2.2. Unobservable student characteristics

A second threat to correctly identifying the causal effect of MTGs on student performance are so-called unobservable student characteristics. One such characteristic is students' innate ability. We cannot directly observe a student's ability. However, not accounting for student's ability may bias the results of statistical analyses if ability jointly determines *both* the independent and dependent variables.

The Idea: For example, a high-ability student might get an A at the MTG as well as an A as the final course grade, while the low-ability student might obtain a C for both his MTG as well as his final course grade. Simply regressing the MTG on the final course grade without accounting for innate ability will result in estimates suggesting that the MTG "caused" the final grade. In other words, due to such unobserved variables such as ability, results from regression analysis will likely be biased upward and overstate the importance of midterm letter grades for subsequent performance. While imperfect proxies for ability exist, we do not have access to data such as SAT or GRE scores, nor would this guarantee that we account for all unobserved characteristics.

In an "ideal" world, we would address this problem using experimental methods. We might randomly assign students to various treatments with MTGs of higher and lower values, while a control group received no grade guidance at all. As assignment to these groups is random, the average ability across the students in these groups should be similar, while only the treatment (i.e. the letter grade received) differs. This would allow us to estimate how receiving higher/lower grade affects—on average—subsequent student performance.

However, for ethical reasons, it is impossible—and we would not wish to—randomly assign grades to students in real courses. Consequently, an alternative methodological approach is needed to approximate the random assignment of students to either low or high grades.

To illustrate the idea behind our approach, assume two students who have the same innate ability, talent, and commitment and exert the same effort—that is to say, they do not differ on these unobservable characteristics. However, the performance of these two students is—for whatever random reason—just ever so slightly different. For example, one student's scores across all assignments in the first part of the semester might yield an average of 89.6, while the other obtains an 89.4. Such a small difference may be due to an exam focusing slightly more on one topic rather than another, or a random event in the student's life that caused them to miss an assignment or focus less on a particular reading. In short, the performance of these two students does not differ significantly. However, due to rounding, the former student with an average score of 89.6 will receive an A while the student with an average of 89.4 will receive a B.

This scenario results in a situation analogous to that of an actual experiment: The two students, who are otherwise completely identical, have been assigned to one of two groups with different treatments. Importantly, this assignment is quasi-random as we have no reason to believe that the student receiving the 89.6 and the student receiving the 89.4

systematically differ in talent, effort, or any other feature that might drive performance in subsequent assignments. Remaining omitted variables can only influence the results when students who are on either side of the threshold systematically differ in these omitted characteristics (Owen 2010). If the only difference between students who are close to either side of the threshold is that one set received a B and the other did not, then we can attribute a causal role to receiving the B in determining future course performance. In other words, comparing similar students, where one received an A rather than a B, allows for examining the causal effect of letter grades while accounting for unobserved differences.

We implement this idea by employing a regression-discontinuity design. In this, we follow Owen (2010), Grant and Green (2013), Main and Ost (2014), and Oettinger (2002) and exploit a common feature in the typical US grading system: the presence of thresholds, which divide a continuous range of scores (ranging from 0 to 100 points) into discrete grade units of A, B, C, D, and F. Given that students just above and just below grade cutoffs do not differ significantly, we conceptualize receiving the higher possible letter grade as a treatment relative to the baseline of the lower available grade. We hypothesize that this difference in grades (without a corresponding difference in student ability, talent, and commitment) affects outcomes.

Operationalization. To implement a regression discontinuity design, we must operationalize key variables within the context of a specific set of students. We conduct our research on an "Introduction to American Government" course that is required for all graduates at a large research university. The class was large, with over 200 students, and covered a typical set of topics ranging from constitutional law, history, and institutions, to public opinion, elections, and political parties. The class met three times per week in the afternoon for 50-minute lecture sessions. Students' activities included three exams, short weekly reading quizzes, as well as attendance and participation during lectures as measured by "clicker" activities.

Our first operationalization, the treatment, was the set of cutoffs in so-called MTGs. MTGs were given to students roughly in the middle of the semester. The MTGs are calculated by incorporating the scores of one exam, several quizzes, and numerous lecture days' worth of "clicker" questions and attendance checks. These midterm grades used the simple A, B, C, D, and F letters based on the common 90/80/70/60 thresholds. No pluses or minuses were used in MTGs. The midterm grades were explained to students in the syllabus, by e-mail, and in class, including the weighting and the relationship between continuous scores and letter grades.[1]

Importantly, the assumption of randomness only applies to students whose scores are very close to the threshold between letter grades. For instance, while we can reasonably assume that two students with 89.6 versus 89.4 do not differ significantly in ability and talent (even though one receives an A and the other a B), we cannot make the same assumption for two students with average scores of 82 and 98 (even though in this case one student also received an A while the other obtained a B). For this reason, regression discontinuity designs focus only on those observations (i.e., students) whose scores are very close to the thresholds. This forces the statistical analysis to focus on only otherwise comparable students, one of which is "just below" the threshold, while the other is "just above." We use methods developed by Imbens and Kalyanaraman (2012) to determine what range of scores is to be considered close enough to the cutoff. Following their approach, our data should be analyzed using a bandwidth of 2.91 points (out of 100). This means that we compare students up to −2.91 points below a grade threshold to those up to +2.91 points above the threshold.[2] Students with scores outside of this range are omitted from the analysis.

Having identified the set of relevant students and determined whether they were assigned into the higher or lower groups, we need to define the dependent variable. We operationalize "subsequent student performance" in three ways. Our first outcome variable is the students' second-half grade (SHG)—combined "clicker," quiz, and exam scores—in the second half of the semester. In this, we calculate each

[1]One limitation of our design is that we cannot know that students actually received the intended treatment. A midterm letter grade was uploaded and available to each student in the university's online grade portal, and this availability was announced to students in multiple ways. It is impossible for us to actually know whether any individual student, or even how many in the aggregate, viewed their midterm grade for this course. We assume that this rate was very high, given the importance students place on grades—but this is not empirically verifiable. Thus, the effects we measure are best understood as Intent to Treat Effects rather than Treatment Effects. We also note that the underlying, continuous percentage grade was also available to students through self-calculation of all grades returned to them, though was not presented to them directly on the University website. Based on experience with this course, we believe it is unlikely that many students were actively calculating their underlying percentage at this point in the course.

[2]The bandwidth of 2.91 is appropriate for main analysis using students' SHG as the dependent variable. Subsequent analyses have slightly different bandwidth depending on the variance of the dependent variable.

student's score only in material that came after the midterm was received but using the same weighting formula for assignments used throughout the course. This variable can be used to differentiate between Hypothesis 1 and 2.

To test Hypothesis 3 about final exam grades, we utilize students' scores on the final exam grade. This exam was conducted during the final exam period after the final lecture of the course at the end of the semester. To test Hypothesis 4, we use the cumulative score of three types of low-stakes assessments administered during the second part of the course: First, students' scores from several multiple-choice reading quizzes to be completed prior to lectures. Second, student performance on "clicker" questions administered during lectures to test recognition of basic concepts from assigned readings. Third, it includes their scores obtained for attendance and participation. The final exam counted with 26.6% toward the final course grade, while low-stakes assessments in the second half of the semester contributed 20% to the final course grade.

Regression discontinuity approaches calculate the size of the effect due to being assigned a higher grade, rather than a lower grade, as follows. First a local linear regression is estimated on each side of the threshold. As noted above, only observations within the bandwidth are included. This implies that one regression is estimated for all observations with scores of 2.91 below the threshold and up to that threshold, and a second regression for all observation with scores of 2.91 above the threshold. Second, each regression then enables the estimation of the value at 0 but informed by the patterns on only one side of the threshold. In other words, the regressions yield two estimates of the outcome at the same value of the threshold. The value from the lower-side of the threshold represents the estimated outcome for the lower-grade group, while the value from the upper-side of the threshold captures the estimated outcome of the higher-grade group. Third, the difference between the two estimates is the estimated Local Average Treatment Effect (LATE). This treatment effect is the estimated coefficient measuring the impact of receiving a higher MTG on subsequent student performance.[3]

[3]We implement this method using statistical packages in STATA (rdrobust) and R (rdd). Methods for estimating uncertainty in these parameters have been developed and summarized by Calonico, Cattaneo, and Titiunik (2014), Imbens and Lemieux (2008), and others.

3. Findings

3.1. Effect of MTGs on subsequent course performance

We first estimate the overall effect of MTGs on subsequent student performance as measured by the above-mentioned students' SHG. The independent (or treatment) variable is whether a student received a "high" (rather than a "low") MTG. The US grade system features four thresholds between each of A, B, C, D, and F.

As a first approach, we create an aggregate measure of "high" relative to "low" MTGs. To aggregate scores across the four thresholds, we rescale each student's MTG to reflect their distance from the nearest grade threshold (89.5, 79.5, 69.5, and 59.5). We recode each threshold to the same value of zero and calculate the distance from that threshold. Consequently, 80.5 and 90.5 have the same value of 1.0, while 79 and 89 each have the same value of –0.5.

We find that receiving the lower grade at the cutoff between two letter grades resulted in an improved performance in the second half of the semester by 6.47 points. For example, a student with an MTG of 89 scored, on average, 95 in the remainder of the semester.

In Figure 1, we illustrate the pattern in the data. The individual dots are a scatter plot of student MTGs (*X*-axis) and SHG based on all assessments from the second half of the semester (*Y*-axis). The two black lines represent localized linear regression curves computed using the observations of students who received lower grades (left of zero) and those who received higher grades (right of zero). In each case, grades on the opposite side of zero are not used in calculating the localized model. The dashed lines around each represent 95% confidence intervals. The gap between the two fit lines at the threshold illustrates the effect size, i.e. the difference of 6.47 points.

Figure 2 visualizes the same data in a different manner. The histogram shows that, on average, students who received the treatment of a lower grade outperform those who received a high grade. For instance, significantly more students scored between 70 and 80 when receiving a higher grade, while a remarkable number of students who received a lower grade obtained scores between 80 and 90.

However, combining thresholds might actually disguise more than it reveals. As with most causes in social science, we expect the effects of a grade intervention to be heterogeneous and conditioned by other attributes. It is possible, for instance, that the strength

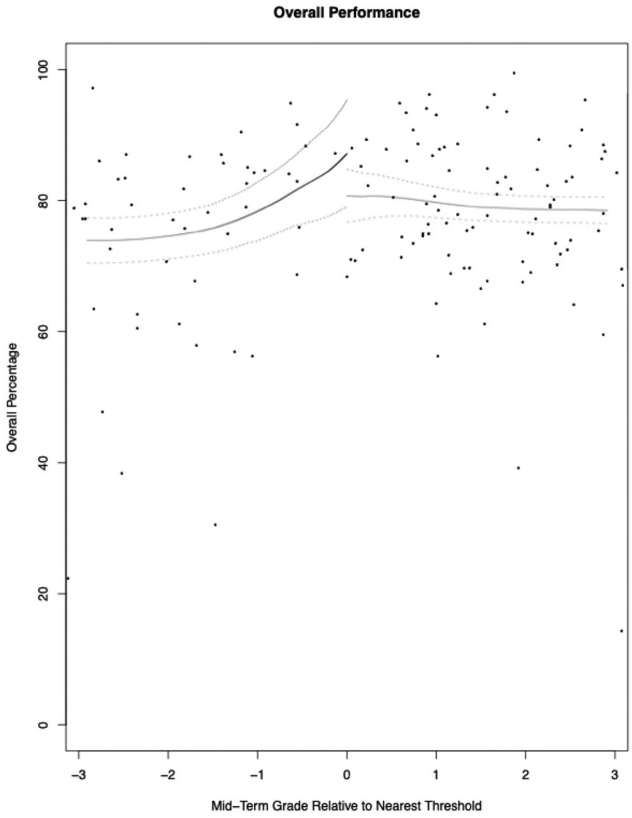

Figure 1. Aggregate effect of midterm grades on the second-half grade.
Note: The graph shows that students who received a midterm grade just below a threshold perform slightly better in the remainder of the semester. For instance, note the absence of observations with a second-half grade below 75 and between −1 and 0, when compared with the range between 0 and +1. This graph combines data from all four thresholds between the letter grades A, B, C, D, and F.

of the effect differs across thresholds. The effect of grades on subsequent student performance may be stronger at the upper and the lower ends of the grade distribution (A/B and C/D thresholds), but not at intermediate grade cutoffs (B/C threshold). A strong student at the border between A and B, given the lower of the two, may work harder to achieve the maximum grade. Similarly, receiving a D might provide a strong incentive to improve performance so as to guarantee receiving credit for the class. In contrast, a student in the middle of the grade distribution, who would require a significant deviation from their existing performance to either earn an A or to fail the course, may be less sensitive to the difference between a B and a C. In fact, Oettinger (2002) argues that grades have a positive effect on subsequent performance primarily at the upper and lower parts of the grade distribution but not at intermediate levels.

To test for this possibility, we conduct separate regression discontinuity models for each threshold. Specifically, we estimate the effect of receiving an A rather than a B, B rather than C, and C rather than D.[4]

Table 1 offers the findings. The results show that students receiving a B instead of an A did about 8.22 points better in the remainder of the semester than those who received an A initially. Similarly, the performance of students who received a D rather than a C increased by about 20.64 percentage points. Both effects are statistically significant at 5% levels. In contrast, receiving a C instead of a B does not have a statistically significant effect on subsequent course performance. As an illustration, these effect sizes imply that a student with a MTG of 89 received (who consequently received a B instead of an A), on average, a grade of 97 for the coursework completed in the second part of the semester. The effect is stronger for students threatened with failing the course: student with a MTG of 69 received (who consequently received a D instead of a C), on average, a grade of 90 for the coursework completed in the second part of the semester.

In sum, both the aggregated analysis across all thresholds as well as the separate estimations for each threshold suggest that, on average, receiving a low MTG results in students improving their performance in the second part of the semester. These findings offer strong evidence in favor of Hypothesis 1, while rejecting Hypothesis 2.

3.2. Do students focus on final exam or low-stakes assessments?

While it is encouraging that students improve their performance after receiving low MTGs, we are interested in understanding *how* students accomplish this. We investigate two possible mechanisms. Hypothesis 3 suggests that students might focus on improving their performance on the final exam, while Hypothesis 4 argues that students will instead re-allocate their attention to the low-stakes assignments throughout the remainder of the semester.

To test the validity of these Hypotheses, we re-estimate the model with different dependent variables instead of using the SHG, which combines the scores

[4]We do not analyze the discontinuity between D and F grades because of an insufficient quantity of students in this grade range. For this test, the Imbens-Kalyanamaran method yields bandwidths that are not substantively defensible. Thus, we fix a bandwidth of 2.5 percentage points. We note that this test is based on small sample sizes within these bandwidths.

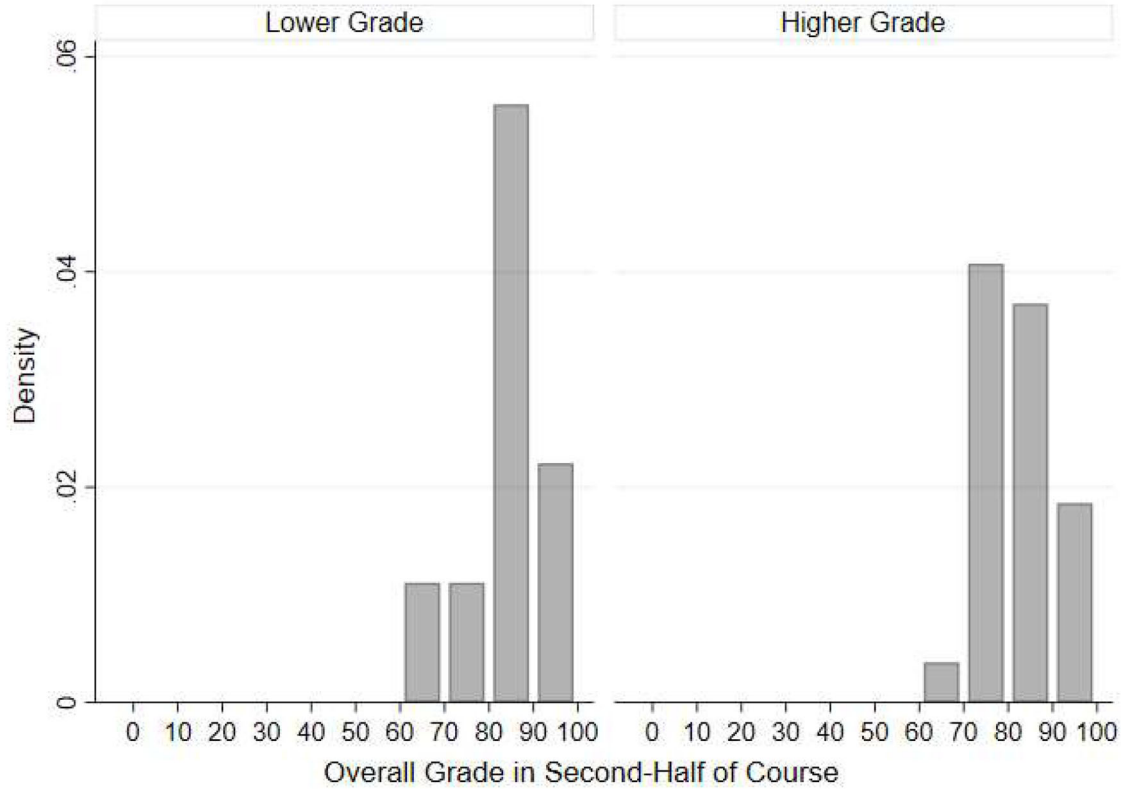

Figure 2. Histogram of second-half grades conditional on treatment.
Note: The histogram compares two grade distributions, one for the set of students who received lower grades while the other represents the grades of students with higher grades. The graph shows that students scoring between 80 and 90 are over-represented among the former, while scores between 70 and 80 are over-represented among the latter.

Table 1. Effect of midterm grades on second-half grade, by grade cutoff.

Threshold	LATE	SE	*N*
A/B	8.22*	3.87	39
B/C	−10.02	6.30	34
C/D	20.64*	5.78	27

Note. The table shows that students who received a midterm grade of B performed 8.22 points better than students just above the threshold who received an A. Similarly, students just below the threshold receiving a D performed better than those just above the threshold obtaining a C. However, letter grades do not have an effect at the intermediate threshold between B and C.
*= $p < 0.05$.

Table 2. Effect of midterm grades on student behavior.

	LATE	SE	*N*
Final Exam	2.88	3.79	153
Low Stakes Assessments	25.90*	11.84	104

Note. The table shows that students who received a grade just below the threshold did not perform better in the final exam than students who received the higher grade. However, they did significantly better with low-stakes assessments in the second half of the semester. On average, their performance was 25.90 points better than those who received the higher grade.
*= $p < 0.05$.

of the final exam with that of the various low-stakes assessments. To test Hypothesis 3, the first analysis estimates the effect of MTGs on the final exam grade only. Conversely, a second analysis estimates the effect of MTGs on low-stakes assessments only to examine Hypothesis 4. The results are summarized in Table 2 as well as Figures 3 and 4.

The numerical results displayed in Table 2 indicate that low MTGs indeed result in distinct student behavior. Specifically, the performance of students on the final exam does not differ significantly across students who received high or low MTGs. However, the performance of these two groups differs significantly with respect to low-stakes assessments such as participation, reading quizzes, and attendance. The treatment effect is large (25.90 percentage points) and statistically significant (p-value of 0.029).

Figure 4 illustrates this pattern in the data. While the difference in the regression lines at zero is indistinguishable for final exam grades, a large gap between the regression lines at zero characterizes the distribution of scores for low-stakes assessments. Figure 4 visualizes the data in the form of Histograms displaying the grade distributions for final exams and low-stakes assessments across the two student groups. On the left are those that fell on the underside of the

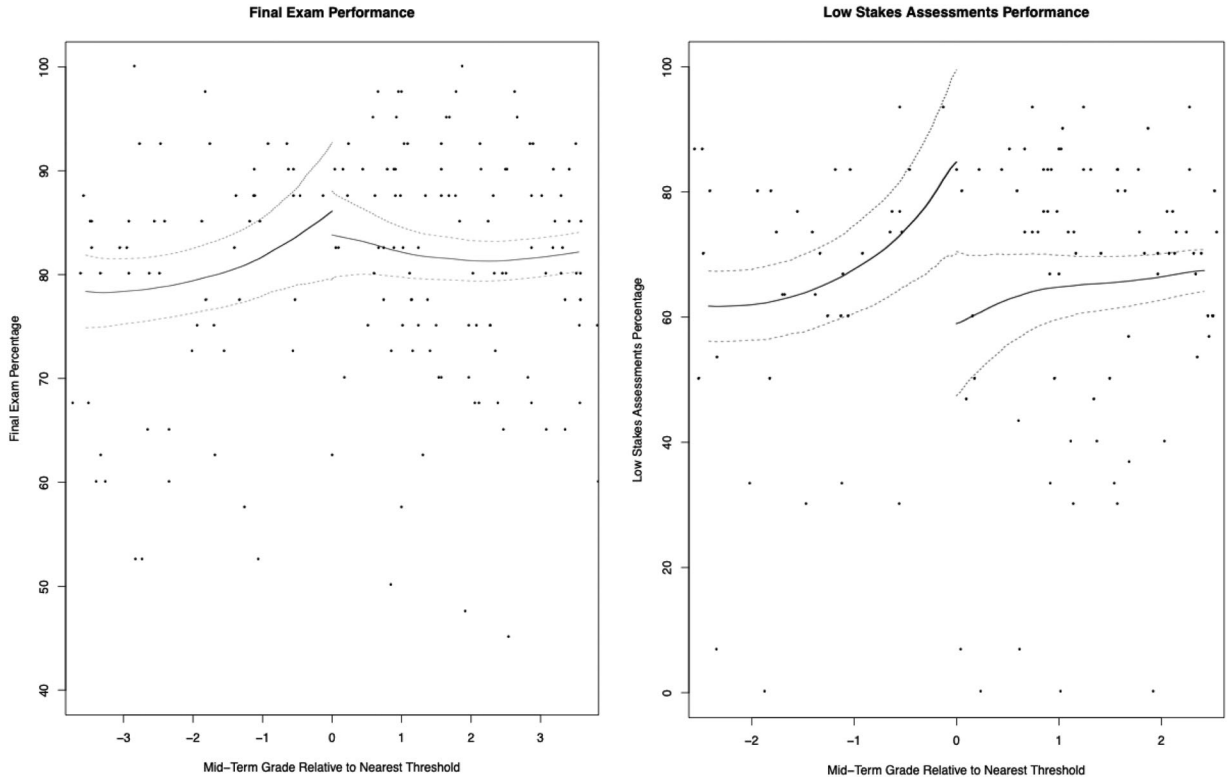

Figure 3. Effect of midterm grades on final exam performance as well as performance on low-stakes assessments.
Note: The figure illustrates that the final exam grades do not differ significantly between students below versus above the threshold. In contrast, students with low MTGs perform significantly better on low-stakes assessments in the second part of the semester, while the grades on these assessments do not improve for students with high MTGs. In particular, a number of students who received high MTGs score between 0 and 10 points on low-stakes assessments in the second part of the semester.

threshold, and thus received a lower MTG. On the right are those who received a high MTG. The grade distribution for the final exam is not significantly different between both groups, though the small number of students with high MTGs who received final exam scores between 50 and 70 explain the point estimate displayed in Table 2. However, clear differences emerge across the groups with high versus low MTGs with respect to scores on low-stakes assessments. We see that those who received a higher midterm grade were more dispersed, with a number of students receiving zeros or scores between the low 30 s and high 60 s. In contrast, the distribution of scores on low-stakes assessments for students receiving the lower MTG features hardy any students scoring less than 70. Tellingly, the number of students obtaining 90 and above is almost six times higher than among students who received the high MTG.

In sum, the analyses show that there is no empirical support for Hypothesis 3. Students who receive a low MTG do not appear to improve their performance on the final exam. Instead, these results offer strong support for Hypothesis 4, suggesting that students achieve improved course performance by focusing on low-stakes assessments such as attendance and participation.

3.3. How does the effect depend on student characteristics?

The preceding analyses show that (a) students who receive lower MTGs exhibit improved performance in the remainder of the semester, and (b) student accomplish this primarily by boosting their performance on low-stakes assessments. In this section, we present a series of additional analyses to examine how student characteristics may condition these findings. Specifically, we explore how the effects may depend on students' major, age, and gender.

Major: Social Science versus nonsocial Science. As noted above, we study the effects of MTGs on subsequent performance in the context of a class required of all students graduating from this university. As a result, the sample includes students majoring in many different fields. We want to explore whether some types of majors respond more strongly to midterm grades than others. Specifically, we are interested in the difference between social science majors (Political

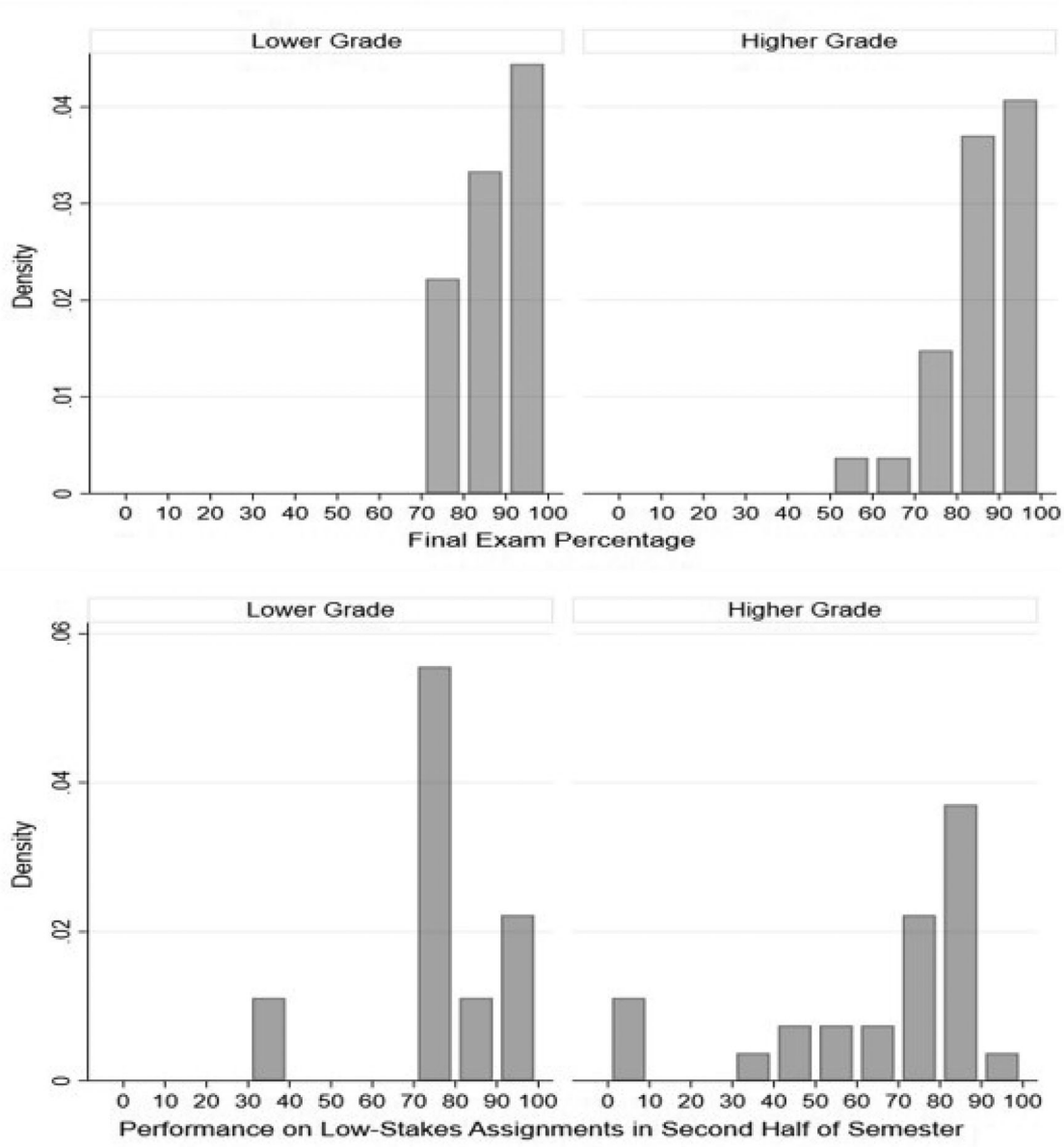

Figure 4. Histogram of low-stakes assessments conditional on treatment.
Note: The histogram compares the distributions of final exam grades and grades on low-stakes assessments across the treatment and control group. The graph shows that the distribution of final exam grades differs little between the groups. In contrast, the distribution does differ with respect to low-stakes assessments, with students who received lower MTGs, on average, score higher than students who obtained higher MTGs.

Science, Sociology, Economics, etc.) and nonsocial science majors (Engineering, STEM, etc.).

It might be the case that social science majors find the content of the course "Introduction to American Government" more directly applicable to their major (Bunte 2019). For this reason, they might have more intrinsic motivation to study for the course irrespective of the MTG received. As a result, social science majors might not respond differently after receiving a high MTG. In contrast, nonsocial science students may find the course less directly applicable to their major and thus may deprioritize the course in response to an early signal of success.

To examine this possibility, we classify the majors represented in our sample into two broad categories: social science students and nonsocial science students. We exclude all students with undeclared majors. Where students have multiple majors, we include them in the category of their first-listed major. We then re-estimate the model examining the effect of MTGs on low-stakes assessments for each of the two sub-populations.

Table 3. Effect of midterm grades on subsequent performance on low-stakes assessments by major, age, and gender.

	LATE	SE	*N*
Nonsocial Science	−30.30○	15.74	64
Social Science	3.52	11.03	47
Freshmen	−25.74○	14.17	72
Older Students	−9.88	23.43	35
Men	−31.18*	14.92	67
Women	4.50	9.20	48

Note. The table shows that the effect of MTGs on subsequent student performance is stronger among students with unrelated (nonsocial science) majors, while the effect of MTGs is insignificant among social science majors. Similarly, the effect exists among freshmen, but not older students. We find the strongest effect for gender, where men strongly respond to grade signals with either increased (if low MTG) or decreased (with high MTG) performance, while women appear to be unaffected.
○= $p < 0.10$.
*= $p < 0.05$.

The results, presented in the top panel of Table 3, provide limited support for these expectations. There is some evidence that nonsocial science students who received higher grades did, in fact, do worse going forward (–30.30 percentage points). However, this difference is not significant by conventional standards ($p = 0.062$). In contrast, the effect of MTGs for social science majors is substantively small and not statistically significant.

Age: Freshmen versus Upper Classmen. We also analyze whether the effect of midterm grades on subsequent scores differs by students' year of study. It may be the case that that first-year students are more sensitive to MTGs than students in later years. MTGs are the first signal that informs first-year students how well they are doing in their new environment. Further, eight weeks into their first semester in college, first-year students are likely still in the process of establishing social bonds with other students, and high grades may help impress fellow students (Selby and Murphy 1992). Conversely, older students have more experience against which to judge the signal of midterm grades and are likely to have already established a social environment that does not rely on grades for validation and thus do not respond to grade incentives (Grove and Wasserman 2006). For this reason, first-year students might be expected to respond more strongly to a low MTGs by improving their performance on low-stakes assessments, while older students do not exhibit such an effect after a low MTG.

To test for this possibility, we re-estimate the model examining the effect of MTGs on low-stakes assessments for each of the two sub-populations. The findings—presented in the center panel of Table 3—provide little support for this expectation. First-year students who received low MTGs did increase their performance on basic course tasks by 25.74 points, but this difference is not statistically significant at conventional thresholds. In contrast, older students did not respond to a low MTG as evidenced by an estimate that is not statistically significant (p-value of 0.692).

Gender: Male versus Female. Lastly, we examine whether responses to MTGs differ by gender. Much existing research suggests that women and men might react differently to grades (Owen 2010; Jensen and Owen 2000). Women have been shown to have a greater awareness of their course performance (Rask and Tiefenthaler 2008). As a result, they tend to exert more continuous effort in the class throughout the semester. Consequently, we would expect that the signal of a low MTG is less important for women—their study habits and class engagement will continue relatively as before. In contrast, we may expect men to respond more strongly to an early signal of success with lower subsequent effort.

The bottom panel of Table 3 presents the results. They show a substantial divergence between men and women.[5] The performance of male students on low-stakes assessments decreases substantially upon receiving a high MTG, but increases significantly when receiving a low MTG. The difference in the grade on low-stakes assessments between men just above and just below the threshold was 31.18 points (p-value of 0.043). In contrast, women performed about the same regardless of whether they received the lower or higher MTG (p-value of 0.640). This may imply that male students were more likely to accept initial signs of success as sufficient while female students maintained their effort levels throughout the course.

4. Practical implications of the research results for instructors

What are practical implications of the research results for instructors? Presumably, instructors would want to utilize the positive effect of low MTGs whereby students are motivated to increase class attendance and participation. Based on our findings, we make several suggestions.

[5]It is possible that the results for Social Science versus non-Social Science as well as for male versus female might not offer distinct findings if gender and major were strongly correlated. However, the descriptive data reveal that Social Science majors are not all female and non-Social Science majors are not all male. Instead, 51% of Social Science majors are female with 49% being male, while 72% of non-Social Science majors are male with 28% being female. The correlation between major and gender equals 0.23.

4.1. The immediate implications of our findings

Our findings suggest that students improve their performance in the second part of the semester by focusing on low-stakes assessments. One recommendation for instructors might be to encourage students to prioritize these assignments over studying for the final, because it essentially guarantees points in return for predictable effort. In contrast, focusing on the final exam is a high risk-high reward strategy. To ameliorate this, instructors could work toward reducing the perceived risk by providing specific guidance for how to study for the final. Existing research suggests that providing nongraded practice tests as well as examples of "good" answers helps reduce students' anxiety (McKeachie and Svinicki 2013, 91).

4.2. The quantitative approach: Use grades as motivation

We outline two strategies that instructors might implement to increase students' motivation in the second half of the semester. The first strategy proposes in-class exercises and grading schemes intended to motivate students by pointing out the grade they could achieve if they were to study hard during the latter part of the semester. Instructors could use the grades themselves as explicit motivational tools to give students with grades just below a threshold hope for how to make it across the threshold, and students with grades just above a threshold with guidance regarding what is required to get to the next letter grade.

A first instructional exercise is inspired by Barkley (2009) who describes an assignment designed to help students see the correlation between their efforts and their exam scores. We apply this exercise to MTGs. In preparation, we recommend creating a grade-calculator, which is essentially an excel file containing the grading policies. Students could enter the number of points achieved on past assignments and enter the points they hope to achieve in future assignments in the course, while the excel formulas automatically calculate the expected letter grade based on the inputs. After the MTGs have been posted, instructors could ask students to look up their grade in the online system and write a short 1-page response paper in class answering the following four questions:

1. Describe your emotional response to your Mid-Term Grade (Surprised? Disappointed? Relieved? Pleased?).
2. Review the grades you have received for all assessments in the first part of the semester. Do you see patterns in your course performance? What types of assessments did they do well on versus not? Where is room for improvement?
3. Use the grade-calculator to predict your final letter grade for the course if you get full points on every assignment in the second part of the semester.
4. Us the insights from your analysis from step 2 to create an action plan for the remainder of the semester. This plan should be specific and concrete (e.g., "I plan to get at least 85 percent of all clicker questions correct.")

A second way to use grades as tools for motivating students would involve grading strictly at the beginning of the semester while becoming more lenient toward the end of the semester. One way to formalize such grading behavior would be to offer dropping the scores of some assignments by the end of the semester. However, the MTGs would be calculated while temporarily ignoring such favorable rules. This results in "worse" grades for students initially, hopefully motivating them to work harder for the remainder of the semester.

A third option involves peer pressure. Based on existing research by Santoro-Ratliff and Bunte (2020), student who learn that their friends received better grades than themselves will improve their subsequent performance. Consistent with the findings of our study where bad grades improve student performance while good grades do not lead to a deterioration of grades, the peer effect identified by Santoro-Ratliff and Bunte also works only in the positive direction: learning that a friend has a lower grade does not negatively affect subsequent performance, while learning that a friend has a higher grade does increase subsequent performance.

4.3. The qualitative approach: Get students thinking about learning, not grades

In contrast to explicitly utilizing grades to motivate students, a second strategy proposes student activities that get students thinking about learning, not grades, and thereby increase students' intrinsic motivation.

This might also involve a reflection paper after receiving their MTG, but with different questions. Instead of focusing how they can achieve the next best grade level, instructors could ask students about their life goals and how this course might allow them

to get there. If the student believes the knowledge is relevant and important then the student is more likely to internalize her learning and work toward building more knowledge, rather than just focusing on the grade (Kuhn and Weinstock 2002). Inspired by Farias, Farias, and Fairfield (2010), we would recommend instructors to ask the following questions:

1. Think about the kind of professional career or even the kind of life you hope to live one day.
2. Following this, what knowledge and skills you are likely to need–but currently do not possess–to live that kind of life. Make a list with at least five skills.
3. Lastly, brainstorm how this class might help you obtain one of the skills on your list/How might you learn these skills in this class during the second part of the semester? How should your study habits change to acquire the skills in the context of the present class?

5. Conclusion

In this article, we analyze how grades affect subsequent student performance. We find that low grades appear to incentivize students to improve their subsequent course performance. A similar pattern is not found among those who received the higher grade. This effect is primarily found among students at the upper (A students who received a B) and lower (C students who received a D) thresholds. We do not find evidence that high grades discourage students, but instead show that lower midterm grades result in improved performance for the remainder of the semester.

In addition, we analyze how students accomplish these improvements by examining two potential mechanisms. We find that students do not appear to increase their performance on final exams, presumably because this is a risky strategy of "saving" the final course grade after a low midterm grade. Instead, students improve their final course grade by improving their performance on low-stakes assessments, such as participation, reading quizzes, and in-class clicker exercises.

Additional analysis reveals that this effect of midterm grades is particularly strong among men: male students show increased effort when receiving lower grades, compared to men receiving higher grades. Women show no such differences based on MTGs received. In addition, we find that this effect is slightly stronger among nonsocial science majors.

We make three contributions to the academic literature. First, while other studies have examined the effect of grades on students' subsequent performance, we unpack this effect by analyzing *how* student behavior changes to accomplish an improvement in grades. Furthermore, we analyze if this effect operates across all students, or whether the effect differs across major, gender, and age of students. Second, we conduct a methodologically rigorous study of how past grades affect future grades. Specifically, our research design addresses potential selection bias which might arise if the sample analyzed is based on students self-selecting into a particular course or major. We avoid this issue by analyzing students enrolled in a course required for all students at a university, irrespective of major. Third, we employ a regression discontinuity methodology. This allows us to accurately identify the causal effect of grades on subsequent student performance.

Our research also has several practical implications for instructors. We outline two strategies that instructors might implement to help students to improve their performance in the second part of the semester. One strategy proposes in-class exercises and grading schemes intended to motivate students by pointing out the grade they could achieve if they were to study hard during the latter part of the semester. The other strategy proposes student activities that get students thinking about learning, not grades, and thereby increase students' intrinsic motivation. Further research is needed to evaluate which of these two strategies is most able to improve students' performance after receiving MTGs.

References

Bar, T., V. Kadiyali, and A. Zussman. 2009. "Grade Information and Grade Inflation: The Cornell Experiment." *Journal of Economic Perspectives* 23 (3): 93–108. doi: 10.1257/jep.23.3.93.

Barkley, E. F. 2009. *Student Engagement Techniques: A Handbook for College Faculty*, 336–9. San Francisco, CA: Jossey-Bass.

Bunte, J. B. 2019. "Why Do Students Enroll in Political Science Classes?" *PS: Political Science & Politics* 52 (2): 353–60. doi: 10.1017/S1049096518002056.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6):2295–326. doi: 10.3982/ECTA11757.

Cameron, J., and W. D. Pierce. 1994. "Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis." *Review of Educational Research* 64 (3):363–423. doi: 10.3102/00346543064003363.

Cameron, J., and W. D. Pierce. 1996. "The Debate about Rewards and Intrinsic Motivation: Protests and

Accusations Do Not Alter the Results." *Review of Educational Research* 66 (1):39–51. doi: 10.3102/00346543066001039.

Deci, E. L., R. Koestner, and R. M. Ryan. 1999. "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin* 125 (6):627–68. doi: 10.1037/0033-2909.125.6.627.

Dev, P. C. 1997. "Intrinsic Motivation and Academic Achievement." *Remedial and Special Education* 18 (1): 12–19. doi: 10.1177/074193259701800104.

Farias, G., C. M. Farias, and K. D. Fairfield. 2010. "Teacher as Judge or Partner: The Dilemma of Grades versus Learning." *Journal of Education for Business* 85 (6): 336–42. doi: 10.1080/08832321003604961.

Grant, D., and W. B. Green. 2013. "Grades as Incentives." *Empirical Economics* 44 (3):1563–92. doi: 10.1007/s00181-012-0578-0.

Grove, W. A., and T. Wasserman. 2006. "Incentives and Student Learning: A Natural Experiment with Economics Problem Sets." *American Economic Review* 96 (2):447–52. doi: 10.1257/000282806777212224.

Guskey, T. R. 2017. "Five Obstacles to Grading Reform." Educational, School, and Counseling Psychology Faculty Publications 6, 1–8. http://www.ascd.org/publications/educational-leadership/nov11/vol69/num03/Five-Obstacles-to-Grading-Reform.aspx.

Haladyna, T. M. 1999. *A Complete Guide to Student Grading*. Boston, MA: Allyn and Bacon.

Hattie, J., and H. Timperley. 2007. "The Power of Feedback." *Review of Educational Research* 77 (1):81–112. doi: 10.3102/003465430298487.

Imbens, G., and T. Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2):615–35. doi: 10.1016/j.jeconom.2007.05.001.

Imbens, G., and K. Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of Economic Studies* 79 (3): 933–59. doi: 10.1093/restud/rdr043.

Jensen, E. J., and A. L. Owen. 2000. "Why Are Women Such Reluctant Economists? Evidence from Liberal Arts Colleges." *American Economic Review* 90 (2):466–70. doi: 10.1257/aer.90.2.466.

Kuhn, D., and M. Weinstock. 2002. "What is Epistemological Thinking and Why Does It Matter?" In *Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing*, edited by B. Hofer and P. Pintrich, 121–44. New York: Routledge.

Main, J. B., and B. Ost. 2014. "The Impact of Letter Grades on Student Effort, Course Selection, and Major Choice: A Regression-Discontinuity Analysis." *The Journal of Economic Education* 45 (1):1–10. doi: 10.1080/00220485.2014.859953.

McKeachie, W., and M. Svinicki. 2013. *McKeachie's Teaching Tips*. Boston, MA: Cengage Learning.

Oettinger, G. S. 2002. "The Effect of Nonlinear Incentives on Performance: Evidence from 'Econ 101'." *Review of Economics and Statistics* 84 (3):509–17. doi: 10.1162/003465302320259501.

Owen, A. L. 2010. "Grades, Gender, and Encouragement: A Regression Discontinuity Analysis." *The Journal of Economic Education* 41 (3):217–34. doi: 10.1080/00220485.2010.486718.

Rask, K., and J. Tiefenthaler. 2008. "The Role of Grade Sensitivity in Explaining the Gender Imbalance in Undergraduate Economics." *Economics of Education Review* 27 (6):676–87. doi: 10.1016/j.econedurev.2007.09.010.

Sabot, R., and J. Wakeman-Linn. 1991. "Grade Inflation and Course Choice." *Journal of Economic Perspectives* 5 (1): 159–70. doi: 10.1257/jep.5.1.159.

Santoro-Ratliff, L., and J. Bunte. 2020. "What did You Get? Peers, Information, and Student Exam Performance." Working Paper.

Selby, D., and S. Murphy. 1992. "Graded or Degraded: Perceptions of Letter-Grading for Mainstreamed Learning-Disabled Students." *British Columbia Journal of Special Education* 16 (1):92–104.

Shanab, M. E., D. Peterson, S. Dargahi, and P. Deroian. 1981. "The Effects of Positive and Negative Verbal Feedback on the Intrinsic Motivation of Male and Female Subjects." *The Journal of Social Psychology* 115 (2): 195–205. doi: 10.1080/00224545.1981.9711659.