Economics of Education Review 56 (2017) 133–140

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

The impact of grade ceilings on student grades and course evaluations: Evidence from a policy change☆

Devon Gorry

Department of Economics and Finance, Huntsman School of Business, Utah State University, 3565 Old Main Hill, Logan, UT 84322, United States

ARTICLE INFO

Article history:
Received 17 February 2016
Revised 28 October 2016
Accepted 16 December 2016
Available online 18 December 2016

JEL:
A22 I23 J24

ABSTRACT

This paper analyzes the effects of a grade ceiling policy on grade distributions and course evaluations. Results show that the effects vary based upon the level of the grade ceiling. A ceiling set at 2.8 decreased overall grade point average (GPA) by reducing the number of As and Bs and increasing the number of lower grades given. This low ceiling also increased the number of withdrawals and significantly lowered course evaluations. A ceiling set at 3.2 decreased overall GPA by reducing the number of As and increasing the number of Bs given, but the effects on course evaluations were smaller in magnitude and insignificant.

1. Introduction

Grade inflation is a growing concern among colleges and universities in the United States.[1] Records of grades since 1960 show a nationwide increase in GPA of approximately 0.1 point per decade (Rojstaczer & Healy, 2010). Given that grades have an upper bound, inflation leads to compression at the top and erodes the value of information provided to students, employers, graduate programs, and others (Kamber & Biggs, 2003). Differences in grading across professors and courses may also distort student decisions about what classes to take. In order to combat grade inflation and level student incentives across courses, some colleges and universities have imposed grade ceiling rules that require professors to keep their average class grades below a certain threshold.

Butcher, McEwan, and Weerapana (2014) evaluate such a policy at an elite liberal arts college. They use a difference-in-differences design across treated and untreated departments to analyze the effects of the grade ceiling. This paper replicates their work by using a difference-in-differences design across treated and untreated professor by course combinations to analyze a grade ceiling policy implemented in the business school of a large state university. The business school implemented a grade ceiling policy in the Spring of 2014 in a quest to avoid grade inflation, ensure fairness across classes, and create rigorous standards. The policy recommended that professors of required business school courses maintain average grades no higher than 2.8 for introductory courses and no higher than 3.2 for intermediate courses. This paper confirms several previous findings in the new setting of a large state university as opposed to a small elite college. It also extends previous findings by analyzing two different levels of grade ceilings rather than one and looking at a broader range of student evaluation measures.

First, this paper evaluates how the grade ceilings affected the distribution of grades. With the 3.2 ceiling, the number of As fell while the number of Bs rose and there was little impact on lower grades or withdrawals. This is similar to the finding in Butcher et al. (2014) where the grade ceiling led to fewer As and more Bs in treated departments. This paper extends these findings by showing that a lower grade ceiling of 2.8 led to fewer As and Bs and an increase in lower grades as well as withdrawals. Next, this paper evaluates how the grade ceilings impacted student evaluations of teachers (SETs).[2] This is important from an administration perspective to the extent that schools care about the satisfaction of their students for retention and enrollment. This is also important from a professor's perspective since SETs are often used to evaluate professors for hiring, tenure, promotion, and pay. This paper finds that the high ceiling is only associated with lower teaching ratings and the decrease is insignificant. The low ceiling, however, is associated with significantly lower ratings across a variety of measures. While the reduction in teaching ratings is largest, other ratings of the course also fall. Butcher et al. (2014) also find that

☆ I would like to thank Jesse Backstrom and Gavin Johnson for excellent research assistance, Guy Ballard for help with collecting the data, seminar participants at Utah State University, Weber State University, and the SEA annual meetings for helpful comments, and anonymous referees. All mistakes are my own.

E-mail addresses: devongorry@gmail.com, devon.gorry@usu.edu

[1] See Eiszler (2002), Johnson (2003), Pressman (2007), and Rojstaczer and Healy (2010) for a discussion of grade inflation.

[2] SETs are an almost universal measurement instrument used to evaluate teaching in higher education (see Becker & Watts, 1999).

http://dx.doi.org/10.1016/j.econedurev.2016.12.006
0272-7757/

a grade ceiling led to lower teaching evaluations, but could not evaluate different SET measures as the evaluations in that setting only rated whether students would recommend a professor.

The findings from this paper can inform schools and professors about the impact of grade ceiling policies. While policies do lower grades and reverse grade compression at the top, they may also lead to worse SETs. These results can also further our understanding of the impact of grades on student evaluations. Several studies have explored whether grades influence SETs and find a positive correlation between grades and teaching evaluations.[3] There are many competing theories to explain this finding, but one prominent theory is that students may "reward" professors who give high grades with better evaluations.[4] In fact, it has been suggested that the increased use of evaluations has attributed to recent grade inflation (Eiszler, 2002; Rojstaczer & Healy, 2010; Stratton, Myers, & King, 1994). While many studies document a positive correlation between giving high grades and receiving high evaluations, few are able to provide causal evidence that grading leniency improves evaluations.[5] This paper and previous work on grade ceilings provide evidence that a policy which exogenously lowers grades leads to worse evaluations for the professors that are impacted.

This paper is organized as follows: Section 2 describes the institutional details and the data, Section 3 lays out the methodology, Section 4 provides the results, and Section 5 concludes.

2. Institutional details and data

The data cover required business school classes at a large state university where enrollment on the main campus is over 13,000.[6] Approximately 1900 of those students are enrolled in the business school. The data span Fall, 2011 through Spring, 2015. This time period is chosen because the current course evaluation system was implemented in the Fall of 2011 and the latest data available at the time of analysis were from Spring of 2015.[7]

Starting in the spring of 2014, the business school implemented a policy that established a *recommended* average grade in required business courses.[8] The policy was motivated by the existence of wide variability in mean grades across different sections of the same course. Thus grade ceilings were set to provide equal grade outcomes across courses. Moreover, it was envisioned that grade ceilings would ensure rigor and better differentiate student achievement. The policy states that grades in required business courses "typically should not exceed a class average of 2.8 in [introductory] courses and 3.2 in [intermediate] courses."[9] It was also recommended that professors include the policy in their syllabi so students would be aware of the standards.

Administrative grade data used in this study cover all required business courses from the study period.[10] The data include average class level GPA[11] as well as a breakdown of the number of each grade given and the number of withdrawals from the class. Thus, in addition to analyzing how the policy affects overall class GPA, this paper assesses how the distribution of grades and the number of withdrawals changed with the grade ceiling.

The grade data is also used to create the treatment and control groups. For each professor by course combination, the pre-policy average GPA is calculated across all of those professor by course classes. If the pre-policy average is above the corresponding grade ceiling, the policy is binding and these professor by course combinations are coded as "treated." If the pre-policy average is below the grade ceiling, then the policy should have no impact on grading, and these professor by course combinations are coded as "untreated."[12] Butcher et al. (2014) also define treatment based on whether pre-policy grades are above the ceiling, but they define treatment at the department level instead of the professor by course level. Observations are dropped if they consist of professor by course combinations that show up only in the pre-policy or post-policy period, but not both.

Evaluation data come from IDEA evaluation reports.[13] These evaluations are conducted electronically during the final three weeks of each term but before the final exam week.[14] While students should have a sense of their final grade, they do not know the exact grade they will receive in the course.[15] These evaluations are voluntary, but students receive several reminders if they do not fill them out. The completion rates in the data average nearly

[3] See Johnson (2003) for a list of papers that look at the relationship between grades and SET scores. In addition, see DeWitte and Rogge (2011) for a review of studies that show significant correlations between grades and SET scores. While most studies show significant positive correlations, Bosshardt and Watts (2001) find a negative correlation and DeCanio (1986) finds an insignificant correlation.

[4] It could also be that students with better professors learn more and the grades reflect that learning. Indeed, some research shows a positive link between teaching effectiveness and student evaluations (Beleche, Fairris & Marks, 2012). However, this link is often weak and other research has shown that effectiveness as measured by performance in follow-on courses is insignificantly or negatively correlated with teaching evaluations (Braga, Paccagnella, & Pellizzari, 2014; Carrell & West, 2010; Weinberg, Hashimoto, & Fleisher, 2009). Student composition and class setting may also explain the positive correlation if preexisting student or class characteristics are associated with both earning higher grades and giving better evaluations.

[5] Weinberg et al. (2009) control for background student and professor characteristics. Isely and Singh (2005) and McPherson (2006) use fixed effects models to account for time invariant instructor and course characteristics. These methods help answer the causal question if there are no unobserved or time invariant endogenous factors left out of the model. Krautmann and Sander (1999) instrument for grades with core and graduate classes. This is a valid instrument if core and graduate classes do not impact evaluations aside from the impact they have on grades. Other papers suggest that grades impact evaluations by testing implications of the various hypotheses (Greenwald & Gillmore, 1997) or showing that there is not a link between grades and learning (Braga et al., 2014).

[6] Regional campuses account for another 14,000 enrollments, but this analysis only covers the main campus. This is because the policy only pertains to the main campus. Moreover, students at the main campus are more representative of a typical full time college student.

[7] A change in the required business courses was also announced in the Spring of 2015 which means that courses that are no longer required may attract a different composition of students.

[8] Required business courses included financial accounting principles (ACCT 2010), managerial accounting principles (ACCT 2020), introduction to economic institutions, history, and principles (ECN 1500), introduction to microeconomics (ECN 2010), global economic institutions (ECN 3400), corporate finance (FIN 3400), legal and ethical environment of business (MGT 2050), managing organizations and people (MGT 3110), fundamentals of marketing (MGT 3500), operations management (MGT 3700), principles of management information systems (MIS 2100), business communications (MIS 3200), and business statistics (STATS 2300).

[9] Introductory courses include 2000-level courses and below while intermediate courses represent 3000-level courses.

[10] In accordance with the IRB, professors were given the option to opt out of the study in which case their data is not included.

[11] The GPA is calculated based on a typical 4 point scale where A is coded as a 4, A- is a 3.66... , down to an F which is coded as 0.

[12] A specification where treatment is specified as the distance between the pre-policy average and the grade ceiling is also analyzed for robustness. The results are qualitatively similar, but outliers in grading have large impacts on magnitudes and significance when the distance measure is used.

[13] IDEA is a nonprofit organization that provides evaluation services to colleges and universities nationwide.

[14] This differs from the timing in Butcher et al. (2014) where evaluations are completed during finals week, but is consistent with the historical use of paper evaluations during the last weeks of class and is similar to timing of other online evaluations found in the literature (see Ellis, Burke, Lomire, and McCormack, 2003; Weinberg et al., 2009; and Braga et al., 2014).

[15] In a survey of students taking an introductory psychology course, Gaultney and Cann (2001) find that 71% of students report that they generally receive the final grade that they expect. Nowell and Alston (2007) also show that the majority (58%) of economics students who completed teacher evaluations during the last week of classes correctly predicted their actual grades and almost all (96%) students were within one grade of their actual grade.

70%.[16,17] The IDEA reports include several measures on which students evaluate their professors and classes. The "teaching rating" has students state their agreement with the following statement: "overall, I rate this instructor as an excellent teacher" on a scale of 1 (definitely false) to 5 (definitely true). This rating is similar to the one analyzed in Butcher et al. (2014) where students rate professors on a four point scale from "strongly recommend" to "do not recommend." For the "course rating," students mark their agreement with "overall, I rate this course as excellent" on the same scale. Students also rate the class on progress on relevant objectives.[18] Students rate each objective on a scale from 1 (no progress) to 5 (exceptional progress) and the "progress rating" is a weighted average of these responses. Finally, the summary score represents a weighted average of the teaching, course, and progress ratings where progress ratings are given double the weight of teaching and course ratings. The score is adjusted to account for students' reported desire to take the course, reported work habits, class size, and measures of course difficulty and student effort not related to the instructor.[19] The summary score is standardized to have a mean of 50 and standard deviation of 10. This is the number that is highlighted as the overall performance measure of a professor on IDEA evaluations.

The data are summarized in Table 1. The top panel shows summary statistics for treated courses before and after the grade ceiling policy was implemented. The bottom panel shows the summary statistics for untreated courses before and after the grade ceiling policy was implemented. From the summary statistics, we can see that the mean class GPA goes down after the policy in treated courses but not for untreated courses. A figure of mean grades over time is also presented in Fig. 1. The figure also shows a fall in grades in the post period for the treated courses while grades stay relatively constant over time for the untreated courses. The figure illustrates that grades began falling before the official policy implementation. The fall coincides with the time when the policy was drafted and announced.[20] Results of the paper are robust to omitting this intermediate time period as well as recoding the post policy period to begin during the time when the policy was discussed. Table 1 also illustrates that evaluation scores fall in treated courses after the policy is implemented, but the untreated courses don't show similar declines. Fig. 1 also plots mean teaching ratings over time for treated and untreated courses. While the treated courses have a fall in ratings, the untreated courses stay relatively constant. Again, the fall begins prior to the implementation of the ceiling, but during the time when the policy was being drafted and announced. A difference-in-differences approach is implemented to analyze whether the patterns in the summary statistics and figures are significant and robust to controls.

Table 1
Summary statistics.

	Treated courses			
	Before policy (n=83)		Afterpolicy (n=64)	
	Mean	St. Dev.	Mean	St. Dev.
Class GPA	3.36	0.28	3.18	0.35
Fraction As	0.53	0.20	0.44	0.19
Fraction Bs	0.37	0.18	0.40	0.15
Fraction Cs	0.07	0.07	0.11	0.10
Fraction Ds	0.01	0.03	0.02	0.03
Fraction Fs	0.01	0.02	0.03	0.04
Fraction withdrawals	0.00	0.01	0.00	0.00
Teaching rating	4.36	0.36	4.15	0.49
Course rating	4.19	0.33	4.08	0.44
Progress rating	4.24	0.30	4.14	0.38
Summary score	55.29	5.05	52.95	6.91
Treated	1.00	0.00	1.00	0.00
Post policy	0.00	0.00	1.00	0.00
Intermediate course	0.42	0.50	0.41	0.50
Class size (100s)	0.67	0.48	0.83	0.57
Spring term	0.34	0.48	0.55	0.50
Summer term	0.13	0.34	0.09	0.29
Fall term	0.53	0.50	0.36	0.48
	Untreated courses			
	Before policy (n=72)		After policy (n=62)	
	Mean	St. Dev.	Mean	St. Dev.
Class GPA	2.72	0.30	2.74	0.33
Fraction As	0.26	0.10	0.27	0.12
Fraction Bs	0.38	0.12	0.38	0.13
Fraction Cs	0.21	0.07	0.22	0.08
Fraction Ds	0.07	0.06	0.07	0.05
Fraction Fs	0.06	0.05	0.06	0.05
Fraction withdrawals	0.02	0.03	0.00	0.00
Teaching rating	4.18	0.47	4.15	0.52
Course rating	3.94	0.40	3.82	0.48
Progress rating	4.16	0.31	4.13	0.35
Summary score	52.86	6.24	52.08	6.88
Treated	0.00	0.00	0.00	0.00
Post policy	0.00	0.00	1.00	0.00
Intermediate course	0.44	0.50	0.42	0.50
Class Size (100s)	0.73	0.60	0.77	0.67
Spring term	0.33	0.47	0.61	0.49
Summer term	0.07	0.26	0.05	0.22
Fall term	0.60	0.49	0.34	0.48

Notes: Means and standard deviations are presented by treatment status and policy period. Treated courses are classes taught by professors in a course with mean class grades above the grade ceiling before the policy. Untreated courses are classes taught by professors whose mean grades in a course were below the grade ceiling before the policy. Before policy represents the period before Spring, 2014 when there was not a grade ceiling in place and after policy represents the period from Spring, 2014 onward when the grade ceiling was in place. For both the before and after policy period there are 38 professor by course combinations in the data with 21 in introductory courses and 17 in intermediate courses. The observation numbers represent the number of class observations.

3. Methodology

The grade ceiling policy sets a maximum grade recommendation. Therefore, only treated courses with professors who grade above the recommended levels were impacted by the policy. Untreated courses with professors who graded below the recommended level before the policy were not impacted by the policy. This provides a control group which allows us to better ensure that any changes in grading or evaluations in the treated classes are due to the policy and not to other factors changing over time. A difference-in-differences estimation is used to compare the change in outcomes for treated classes before and after the policy to the change in outcomes for untreated classes. If factors other than the policy that affect grading and evaluations do not change over the time period or if they change similarly for treated and

[16] Analysis of whether the response rate changes for treated courses post policy suggests that there is no significant change overall or for introductory courses. There is a marginally significant decrease in the response rate for intermediate treated courses post policy.

[17] This is lower than the response rate in Butcher et al. (2014) as the school they look at has a policy which requires students to respond; however, it is on the high end of other electronic response rates seen in the literature (see Beleche et al., 2012).

[18] The objectives are chosen by the professor from a list of 12 different options. It is recommended that professors choose 3 to 5. The objectives can be marked as "essential" or "important" where essential objectives get twice the weight as important objectives.

[19] See Hoyt and Lee (2002) for details on the adjustment process. Scores are adjusted only if the adjusted score is higher than the unadjusted score. Results of the analysis remain similar in significance and magnitude if unadjusted summary scores are used.

[20] The policy was drafted and discussed during Summer, 2013, and the final version was announced at the beginning of Fall, 2013.

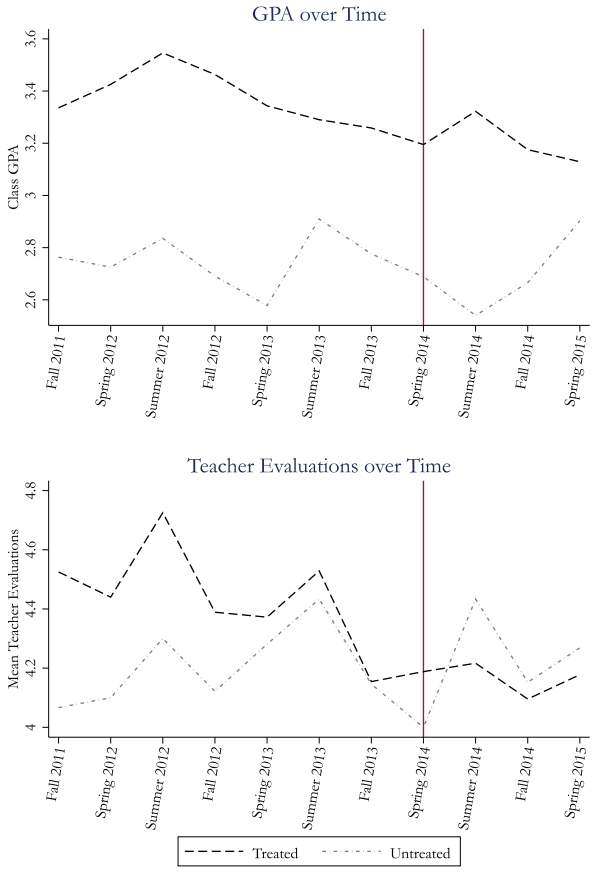

Fig. 1. Average GPA and teaching evaluations over time.

untreated classes, then this approach will estimate the effect of the policy on grading and evaluations.

To implement the difference-in-differences model, the following equation is estimated:

$$Y_{ist} = \beta_0 + \beta_1 Treated_s + \beta_2 PostPolicy_t + \tau Treated_s * PostPolicy_t + X_{ist}\gamma + \eta_s + \nu_t + \epsilon_{ist}$$

where Y represents either grades or evaluations scores, with i indexing the class, s the professor by course combination, and t the semester and year. *Treated* is a dummy equal to 1 for classes impacted by the policy. *PostPolicy* is a dummy for the policy period which includes Spring 2014 and after. X_{ist} represents class level controls that include class size and class size squared. Professor by course fixed effects, η_s, are included to account for time invariant differences that exist across professors in a particular course. Semester fixed effects, ν_t, are included to account for general nonlinear time trends.[21] Note that the professor by course fixed effects and semester fixed effects absorb the effects of *Treated* and *PostPolicy*. This is not a concern because τ is the coefficient of interest and represents how grades or evaluations change for the treated group after the policy relative to the untreated group. All specifications use robust standard errors and are clustered by professor by course.

As an extension, this paper conducts an instrumental variable analysis on treated classes to consider how class GPA affects teaching evaluations where the policy is used as an instrument for class GPA. In this analysis, the first stage is:

$$ClassGPA_{ist} = \alpha_0 + \alpha_1 PostPolicy_t + \alpha_2 X_{ist} + \nu_s + e_{ist}$$

where *PostPolicy* acts as the instrument, X_{ist} now includes additional dummies for spring and fall semester as semester fixed effects cannot be included or they would absorb the policy effect. The second stage is given by:

$$Y_{ist} = \beta_0 + \beta_1 ClassGPA_{ist} + \beta_2 X_{ist} + \eta_s + \epsilon_{ist}$$

This procedure will estimate the magnitude of the effect of grades on teaching evaluations. Results will be consistent as long as the policy period only affects outcomes through its effect on grades.[22]

4. Results

4.1. The effect of the policy on grades

Table 2 presents the effect of the grade ceiling on grades. The first column reports coefficients from the difference-in-differences model where class GPA is the dependent variable. The dependent variables in the remaining columns show the fraction of each particular letter grade given in a class as well as the fraction of withdrawals. The first panel presents the results for all courses combined while the next two panels show the results separated by introductory and intermediate courses. The results are separated across division because different grade ceilings were imposed across division, with a 2.8 ceiling set for introductory courses and a 3.2 ceiling set for intermediate courses. *Treated* and *PostPolicy* coefficients are not presented because these are absorbed by the professor by course and semester fixed effects.

Since the policy puts a ceiling on grades, we expect a decrease in class GPA for treated classes which previously gave grades above the ceiling relative to untreated classes where grades already complied with the policy. Column 1 shows the policy did lead to a decrease in class GPA by 0.132 points for treated classes. When broken out by introductory and intermediate courses, the results are similar with a decrease in GPA of 0.155 points in introductory courses and 0.132 points in intermediate courses. This is in a similar range to the 0.17 point decrease found in Butcher et al. (2014). The remaining grade columns illuminate how faculty complied with the policy. Looking at the top panel for all courses, we see that professors gave significantly fewer As. There were insignificant increases in Bs, Cs, and Fs. Moreover, the fraction of withdrawals increased for the treated classes relative to the untreated classes post policy. This was driven by a relative decrease in withdrawals for untreated classes which could have been caused by the fact that the policy made it more difficult for students to shop for lenient classes. Looking at the effects across introductory and intermediate courses separately suggests that professors responded differently based upon the level of grade ceiling imposed.

In introductory courses, professors met the ceiling of 2.8 by giving fewer As and Bs and substituting with significantly more Cs and a meaningful but insignificant increase in Fs. The policy also led to more withdrawals in treated classes relative to untreated classes. These results differ from Butcher et al. (2014) who find only an increase in the incidence of As and a decrease in Bs with no changes in lower grades or withdrawals. In intermediate courses, professors met the ceiling of 3.2 by giving fewer As and more Bs. This was largely driven by a decrease in straight As and an increase in B minuses. The remaining coefficients on grades and withdrawals are both insignificant and small in magnitude. These

[21] The results are also robust to including linear or quadratic time trends in place of fixed effects.

[22] One way this assumption could be violated is if professors change the way they teach in treated classes upon the policy change. The key identifying assumption needed to attribute the change in evaluation outcomes solely to grades is that teachers respond to the policy by simply shifting their grading cutoffs to adhere to the ceiling. This restriction could also be violated if there are unobserved time effects which impact both grades and evaluations. However, the post policy period does not have any impact on grades or evaluations for the untreated classes which provides some evidence against general time effects.

Table 2
Impact of grade ceiling on class grades.

	Class GPA (1)	As (2)	Bs (3)	Cs (4)	Ds (5)	Fs (6)	Withdrawals (7)
Treated*Post policy	−0.132**	−0.066***	0.011	0.022	−0.003	0.014	0.018***
	(0.057)	(0.023)	(0.024)	(0.015)	(0.010)	(0.011)	(0.006)
Class size (100s)	−0.286**	−0.074*	−0.055	0.080**	0.016	0.032	−0.003
	(0.118)	(0.043)	(0.054)	(0.035)	(0.022)	(0.020)	(0.009)
Class size Sq.	0.084**	0.024*	0.016	−0.025**	−0.004	−0.009	0.000
	(0.039)	(0.014)	(0.018)	(0.012)	(0.008)	(0.005)	(0.003)
Professor by course fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	281	281	281	281	281	281	281
R^2	0.146	0.111	0.048	0.105	0.035	0.063	0.357
Introductory courses							
Treated*Post policy	−0.155*	−0.058**	−0.027	0.035*	−0.006	0.024	0.025***
	(0.079)	(0.025)	(0.027)	(0.019)	(0.016)	(0.016)	(0.007)
Class size (100s)	−0.354*	−0.114*	−0.019	0.081	0.006	0.043	0.002
	(0.192)	(0.057)	(0.071)	(0.055)	(0.031)	(0.032)	(0.013)
Class size Sq.	0.099*	0.031*	0.009	−0.023	−0.001	−0.013	−0.001
	(0.055)	(0.017)	(0.023)	(0.016)	(0.010)	(0.008)	(0.004)
Professor by Course fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	162	162	162	162	162	162	162
R^2	0.200	0.133	0.050	0.139	0.090	0.111	0.441
Intermediate courses							
Treated*Post policy	−0.132	−0.089**	0.072**	0.011	0.005	−0.000	0.005
	(0.076)	(0.040)	(0.034)	(0.022)	(0.010)	(0.009)	(0.004)
Class size (100s)	−0.240***	−0.109	−0.074	0.136**	0.011	0.018	0.014
	(0.076)	(0.088)	(0.141)	(0.061)	(0.028)	(0.024)	(0.017)
Class size Sq.	0.078*	0.050	0.018	−0.053*	−0.002	−0.004	−0.007
	(0.039)	(0.042)	(0.064)	(0.027)	(0.011)	(0.011)	(0.008)
Professor by Course fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	119	119	119	119	119	119	119
R^2	0.141	0.167	0.123	0.125	0.036	0.050	0.289

Notes: GPA is on a 4 point scale. The letter grades and withdrawals represent the fraction of students that received A's, B's, C's, etc. Introductory courses include 2000-level courses and below while intermediate courses represent 3000-level courses. Note that the professor by course fixed effects and semester fixed effects absorb the effects of *Treated* and *PostPolicy*. Robust standard errors clustered by professor by course are in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

findings are more in line with Butcher et al. (2014) who analyze a similar grade ceiling level of 3.33. These results highlight that the effects of grade ceilings depend on the level of the grade ceiling imposed. The higher ceiling led to decompression of grades at the top of the grading scale while the lower ceiling redistributed grades further down the grading scale.

4.2. The effect of the policy on evaluations

Table 3 presents the effect of the grade ceiling on course evaluations. The evaluations used for this paper allow analysis on a number of different rating measures. Each column of Table 3 represents a different evaluation category as described in Section 2. The overall results show that the grade ceiling policy is associated with a statistically significant decrease in teaching ratings by 0.150 points on a 5 point scale. This corresponds to about a quarter of a standard deviation on the teaching rating scale and is similar in magnitude to the 0.111 fall in student recommendations on the 4 point scale found in Butcher et al. (2014). For the combined courses, there are not significant changes in the other measures of student evaluations. The coefficients on the progress rating and the summary score are negative, but statistically insignificant.

The second panel of Table 3 presents the results for introductory courses, where the ceiling was set lower and grades were shifted relatively further down the grade distribution. Here, the negative impacts of the policy show up across most evaluation measures. Not only is teaching rated lower, but the progress on relevant objectives has a significant and relatively large fall and the course rating has a meaningful, but insignificant, fall. In turn, the summary score measure also falls significantly. Thus, students did not just rate the professor more poorly as we saw in Butcher et al. (2014) and in the overall results, but they also rated the course as a whole as well as learning on relevant objectives lower upon policy implementation. The course rating appears to suffer least from the policy and may be more robust to grade changes relative to other measures.

Finally, the third panel presents results for the intermediate courses, where the ceiling was set higher and grades only shifted from As to Bs. Here, there are no significant effects on any of the evaluation measures. However, the magnitude of the effect on the teaching rating is both negative and meaningful. Thus, it appears that only teaching, if anything, was rated relatively worse when the high grade ceiling was imposed. Overall, it appears that the teaching rating may be most sensitive to the grade ceiling.

4.3. The effect of grades on evaluations

As an extension to the difference-in-differences analysis, this paper uses the grade ceiling policy as an instrument to evaluate the effect of grades on evaluations. Table 4 presents the instrumental variable results from regressions on the treated classes. These estimates suggest that a 1 point increase in GPA leads to a 1.530 point increase in the teaching rating, a 0.830 point increase in course rating, a 0.872 point increase in the rating of progress on relevant objectives, and an 18.104 point increase in the summary

Table 3
Impact of grade ceiling on student evaluations of professors.

	Teaching rating (1)	Course rating (2)	Progress rating (3)	Summary score (4)
Treated*Post policy	−0.150**	0.013	−0.064	−1.374
	(0.065)	(0.079)	(0.058)	(1.048)
Class size	−0.435*	−0.319*	−0.344**	−5.947**
	(0.229)	(0.188)	(0.156)	(2.761)
Class size Sq.	0.107	0.080	0.100*	1.725*
	(0.078)	(0.066)	(0.051)	(0.939)
Professor by course fixed effects	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes
Observations	281	281	281	281
R^2	0.128	0.097	0.081	0.084
Introductory courses				
Treated*Post policy	−0.185**	−0.086	−0.156*	−2.750*
	(0.087)	(0.074)	(0.081)	(1.366)
Class size	−0.645*	−0.486*	−0.435*	−7.857*
	(0.310)	(0.280)	(0.250)	(4.365)
Class size Sq.	0.153	0.121	0.117	2.117
	(0.096)	(0.088)	(0.076)	(1.338)
Professor by course fixed effects	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes
Observations	162	162	162	162
R^2	0.212	0.163	0.153	0.148
Intermediate courses				
Treated*Post policy	−0.084	0.193	0.030	0.642
	(0.068)	(0.123)	(0.073)	(1.290)
Class size	−0.334	−0.437	−0.542*	−7.082
	(0.335)	(0.332)	(0.265)	(4.122)
Class size Sq.	0.146	0.180	0.248*	2.986
	(0.153)	(0.147)	(0.122)	(1.837)
Professor by course fixed effects	Yes	Yes	Yes	Yes
Semester fixed effects	Yes	Yes	Yes	Yes
Observations	119	119	119	119
R^2	0.120	0.117	0.110	0.089

Notes: Teaching, Course, and Progress Ratings are all on a 5 point scale where 1 represents a poor rating and 5 represents an excellent rating. The Summary Score measure is on a scale with a standardized mean of 50 and standard deviation of 10. Introductory courses include 2000−level courses and below while intermediate courses represent 3000-level courses. Note that the professor by course fixed effects and semester fixed effects absorb the effects of *Treated* and *PostPolicy*. Robust standard errors clustered by professor by course are in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

score.[23] While a 1 point increase in GPA is large, it is well within the variation observed across class grade averages. The first stage statistics suggest that the instrument is marginally weak in this case (see Stock & Yogo, 2005). However, using weak instrument robust tests, Table 4 shows that the Anderson–Rubin statistics still indicate significant effects of class GPA on the evaluation measures in all cases.

The effects in introductory courses are similar to the overall effects. While the effects in intermediate courses appear large, they are not significant in any category and suffer more greatly from weak instruments.[24] Overall, grading behavior may have large impacts on course evaluations.

5. Discussion

With ever increasing grade inflation, we are likely to see more grade ceiling policies imposed. This paper provides insight into how these ceilings will impact the distribution of student grades. While higher ceilings ease the bunching of grades only at the top of the distribution, lower ceilings can further spread the distribution into the C range and below. The change in withdrawals suggests that lower ceilings may also prevent "shopping" for more lenient classes. These grade ceiling policies also provide a useful identification tool to evaluate the impact of grades on student evaluations.

This paper shows that grades matter for teaching evaluations. Not only do treated professors receive worse teaching evaluations upon implementation of a grade ceiling policy, but the ratings of the courses and progress on objectives also fall when the ceiling is low. The differences across grade ceilings suggest that grade changes may have varying effects on evaluations. While the fall in GPA was similar across introductory and intermediate courses, the shift in introductory courses was driven by significantly more Cs and possibly more Fs. In turn, there were stronger negative effects on a variety of evaluation ratings in introductory courses. Thus, students may be particularly sensitive to receiving low grades when rating their professors. It may also be that the marginal student who is at risk of failing is more sensitive to grade changes when evaluating classes compared to the students who are on the border of receiving an A or a B. The use of student evaluations is ubiquitous and can have large stakes on the careers of professors. This paper adds to the evidence that we should interpret these

[23] This is supportive of the finding in Ellis et al. (2003) that grades are more strongly correlated to teaching ratings than course ratings.

[24] In addition, underidentification cannot be rejected for the intermediate courses. It can be rejected overall and for introductory courses at the 5 percent level or below.

Table 4
Effect of grades on student evaluations of professors.

	Teaching rating (1)	Course rating (2)	Progress rating (3)	Summary score (4)
Class GPA	1.530**	0.830*	0.872	18.104**
	(0.663)	(0.489)	(0.558)	(9.143)
Class size	0.086	−0.221	−0.069	−0.375
	(0.186)	(0.191)	(0.219)	(2.732)
Class size Sq.	−0.028	0.076	0.044	0.121
	(0.076)	(0.073)	(0.079)	(1.117)
Professor by course fixed effects	Yes	Yes	Yes	Yes
Term dummies	Yes	Yes	Yes	Yes
Observations	147	147	147	147
First stage Craig–Donald F Stat	10.406	10.406	10.406	10.406
First stage Kleibergen–Paap F Stat	8.072	8.072	8.072	8.072
Anderson-Rubin p-value	0.011	0.099	0.095	0.0439
Introductory courses				
Class GPA	1.214**	0.583*	0.831	15.142**
	(0.495)	(0.349)	(0.592)	(7.482)
Class size	−0.097	−0.348	0.132	0.272
	(0.293)	(0.309)	(0.411)	(5.170)
Class size Sq.	0.014	0.094	−0.024	−0.213
	(0.078)	(0.092)	(0.121)	(1.494)
Professor by course fixed effects	Yes	Yes	Yes	Yes
Term dummies	Yes	Yes	Yes	Yes
Observations	86	86	86	86
First stage Craig–Donald F Stat	9.492	9.492	9.492	9.492
First stage Kleibergen–Paap F Stat	5.524	5.524	5.524	5.524
Anderson-Rubin p-value	0.017	0.185	0.165	0.083
Intermediate courses				
Class GPA	2.158	1.232	1.081	24.523
	(2.454)	(1.980)	(1.580)	(33.867)
Class size	0.444	−0.054	−0.253	0.818
	(0.784)	(0.511)	(0.397)	(9.159)
Class size Sq.	−0.165	0.028	0.130	−0.211
	(0.359)	(0.227)	(0.174)	(4.122)
Professor by course fixed effects	Yes	Yes	Yes	Yes
Term dummies	Yes	Yes	Yes	Yes
Observations	61	61	61	61
First stage Craig–Donald F Stat	2.011	2.011	2.011	2.011
First stage Kleibergen–Paap F Stat	2.305	2.305	2.305	2.305
Anderson Rubin p-value	0.255	0.471	0.385	0.354

Notes: Teaching, Course, and Progress Ratings are all on a 5 point scale where 1 represents a poor rating and 5 represents an excellent rating. The Summary Score measure is on a scale with a standardized mean of 50 and standard deviation of 10. Introductory courses include 2000-level courses and below while intermediate courses represent 3000-level courses. Robust standard errors clustered by professor by course are in parentheses. The Anderson-Rubin p-value is a weak instrument robust test on the null of a Class GPA coefficient of 0.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

evaluations with caution and take into account grading behavior when evaluating professors.

References

Becker, W. E., & Watts, M. (1999). How departments of economics evaluate teaching. *The American Economic Review, Papers and Proceedings, 89*(2), 344–349.

Beleche, T., Fairris, D., & Marks, M. (2012). Do course evaluations truly reflect student learning? evidence from an objectively graded post-test. *Economics of Education Review, 31*, 709–719.

Bosshardt, W., & Watts, M. (2001). Comparing student and instructor evaluations of teaching. *The Journal of Economic Education, 32*(1), 3–17.

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*, 71–88.

Butcher, K. F., McEwan, P. J., & Weerapana, A. (2014). The effects of an anti–grade-inflation policy at wellesley college. *Journal of Economic Perspectives, 28*(3), 189–204.

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy, 118*(3), 409–432.

DeCanio, S. J. (1986). Student evaluations of teaching: A multinomial logit approach. *The Journal of Economic Education, 17*(3), 165–176.

DeWitte, K., & Rogge, N. (2011). Accounting for exogenous influences in performance evaluations of teachers. *Economics of Education Review, 30*, 641–653.

Eiszler, C. F. (2002). College Students' evaluations of teaching and grade inflation. *Research in Higher Education, 43*(4), 483–501.

Ellis, L., Burke, D. M., Lomire, P., & McCormack, D. R. (2003). Student grades and average ratings of instructional quality: The need for adjustment. *The Journal of Educational Research, 97*(1), 35–40.

Gaultney, J. F., & Cann, A. (2001). Grade expectations. *Teaching of Psychology, 28*(2), 84–87.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading lenience is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209–1217.

Hoyt, D. P., & Lee, E.-J. (2002). Technical report number 12: Basic data for the revised IDEA system.. *The Individual Development and Educational Assessment Center*.

Isely, P., & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *The Journal of Economic Education, 36*(1), 29–42.

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. Springer-Verlag New York, Inc.

Kamber, R., & Biggs, M. (2003). Grade inflation: Metaphor and reality. *The Journal of Education, 184*(1), 31–37.

Krautmann, A. C., & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review, 18*, 59–63.

McPherson, M. A. (2006). Determinants of how students evaluate teachers. *The Journal of Economic Education, 37*(1), 3–20.

Nowell, C., & Alston, R. M. (2007). I thought i got an A! overconfidence across the economics curriculum. *The Journal of Economic Education, 38*(2), 131–142.

Pressman, S. (2007). The economics of grade inflation. *Challence, 50*(5), 93–102.

Rojstaczer, S., & Healy, C. (2010). Grading in american colleges and universities. *Teachers College Record*.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. Andrews, & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of thomas j. rothenberg* (pp. 80–108). Cambridge University Press.

Stratton, R. W., Myers, S. C., & King, R. H. (1994). Faculty behavior, grades, and student evaluations. *The Journal of Economic Education, 25*(1), 5–15.

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *The Journal of Economic Education, 40*(3), 227–261.