

Utah State University

DigitalCommons@USU

---

All Graduate Plan B and other Reports

Graduate Studies


---

5-2021

## Exploiting a Grading Policy Shift as an Instrument to Estimate Impact of Grading on Teacher Evaluations

Gavin Johnson  
*Utah State University*

Follow this and additional works at: <https://digitalcommons.usu.edu/gradreports>

 Part of the [Econometrics Commons](#), and the [Education Economics Commons](#)

---

### Recommended Citation

Johnson, Gavin, "Exploiting a Grading Policy Shift as an Instrument to Estimate Impact of Grading on Teacher Evaluations" (2021). *All Graduate Plan B and other Reports*. 1539.

<https://digitalcommons.usu.edu/gradreports/1539>

This Report is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Plan B and other Reports by an authorized administrator of DigitalCommons@USU. For more information, please contact [digitalcommons@usu.edu](mailto:digitalcommons@usu.edu).



Exploiting a Grading Policy Shift as an Instrument to Estimate  
Impact of Grading on Teacher Evaluations

by

Gavin Richard Johnson

A thesis submitted in partial fulfillment  
of the requirements for the degree

of

MASTER OF SCIENCE

in

Economics

Approved:

---

Ben Blau, Ph.D.  
Major Professor

UTAH STATE UNIVERSITY  
Logan, Utah

2021

Copyright © Gavin Johnson 2021

All Rights Reserved

## ABSTRACT

Exploiting a Grading Policy Shift as an Instrument to Estimate Impact of Grading on  
Teacher Evaluations

by

Gavin R. Johnson

Utah State University, 2021

Major Professor: Ben Blau, Ph.D.  
Department: Economics and Finance

Professors at a university plausibly have an incentive to give higher grades to students, and these higher grades will be reflected in student evaluations, which are used to assess teaching quality, which could have career impacts. This paper takes advantage of a policy shift at the business school at Utah State University that introduced suggested caps on the average course grades that teachers gave. This allowed instrumental variable analysis to correct for bias in OLS estimations of these impacts. The correlation between grades and students' evaluations of teachers was found to be positive suggesting that student evaluations of teachers are biased by the grades that teachers give, making them less useful as a guiding metric.

(24 pages)

## ACKNOWLEDGMENTS

I would like to thank my wife, my parents, other members of my family, and all others who have helped and supported me in this endeavor and through challenges over the last several years.

Gavin Richard Johnson

## TABLE OF CONTENTS

	Page
ABSTRACT.....	2
ACKNOWLEDGMENTS .....	3
List of Tables .....	5
List of Figures .....	5
Chapter	
Introduction .....	6
1. Related Literature .....	6
2. Description of the Data .....	9
3. Empirical Methods .....	11
4. Empirical Results .....	12
5. Robustness Tests .....	13
6. Conclusion and Discussion .....	14
References.....	17
Appendix: Tables and Figures.....	18

## LIST OF TABLES

	Page
Table 1: Selected Summary Statistics .....	18
Table 2: Courses falling within grade policy window, by division.....	18
Table 3: OLS results.....	18
Table 4: 2SLS Teacher Quality Result .....	19
Table 5: 2SLS Course Evaluation Results .....	19
Table 6: 2SLS Progress on Fundamental Indicators Evaluation Results .....	20
Table 7: 2SLS Summary Score Evaluation Results .....	20
Table 8: First stage results.....	21

## LIST OF FIGURES

	Page
Figure 1: Upper division distribution of grades pre-implementation of policy .....	22
Figure 2: Upper division distribution of grades post-implementation of policy .....	22
Figure 3: Lower division distribution of grades pre-implementation of policy.....	23
Figure 4: Lower division distribution of grades post-implementation of policy .....	23

## **Introduction**

Student evaluations of teachers are common in higher education and are increasingly prevalent. They are used to assess teachers and courses and can have career impacts. Thus, teachers have incentive to inflate grades to buy better evaluations. Significant positive correlations have been noted since the 1970s. Theoretical research has suggested the effect of the incentive and other reasons for grade inflation. However, this process involves endogenous effects. Grades could hypothetically impact the evaluations students give teachers, and wanting better evaluations, teachers could give better grades. Therefore, estimates of how grades are correlated to teacher evaluations could be biased, and might not be anything like we expect.

Utah State University introduced a policy at the beginning of 2014 to cap the average grades of individual classes. This policy is a general policy, and so should be exogenous to both student evaluations and the grades that teachers assign prior to the policy change. The change should lower the average grades of classes that would otherwise have grades above the cap. This study uses this policy to estimate the correlation between grades and student teacher evaluations using an instrumental variable approach, and attempts to find if grades do affect evaluations, and if so, to what extent.

### **1. Related Literature**

The general consensus in the literature is that higher grades lead to better teacher evaluations. However, this is not necessarily present or statistically significant once controls are included. Centra (2003), and Marsh and Roche (2000) found no significant effect of grades once other controls were included, with the regression coefficient of grades on evaluations fluctuating around zero. The early research by Holmes (1972) and Feldman



(1976) focused on the correlation between grades and evaluations, with some speculation as to why this might be, while later research attempted to better quantify this result and other factors that could cause this. Feldman (1976) found that any correlation was small, but not necessarily representing a causal relationship. Holmes (1972) found that it was rather the discrepancy between what grade a student thought they deserved and what they received that explained evaluations better. Many possible factors exogenous to teacher quality have been identified that could impact evaluations. Dewitte and Rogge (2011) noted that ideally, these effects should be discounted for the purpose of actually evaluating the teacher. Exogenous variables such as class size or how many classes students have taken could impact evaluations.

One of the factors that has not been directly studied but has been noted in much of the literature is that the evaluations can occur either before or after the grade for the course is known by the student. While this largely depends on the institution, some instructors communicate the predicted grades more clearly to students. This factor can lead to either the expectations impacting the evaluations if the evaluations are conducted before the grades are known, or confirmation effects if the grades received are higher or lower than expected and then the evaluation is subsequently conducted (Feldman, 1976; Holmes, 1972). Teaching style, and not just quality, can impact evaluations. Stapleton addresses this, and finds impacts even after controlling for other observables. The correlations observed between grades and evaluations are different for different teachers, so possibly teaching style impacts both.

Other unobservable factors can also impact evaluations: it has been noted that “teacher productivity”, course load, learning, and other unobservables, can bias estimates

of how grades impact evaluations (see, for example, Ewing, 2012; Braga et al., 2014; and De Witte and Rogge, 2011). This is due to the omitted variable bias that can occur, and possible interactions between variables.

Multiple researchers, such as Love et al. (2010) and Eiszler (2002), have created theoretical models and reasons behind motivations for grading and evaluations. It is noted that there is a general incentive for teachers to inflate grades, as being liked as a teacher can be a part of leading to tenure, as students may 'award' the teachers with better evaluations.

This can also lead to endogeneity problems. Teachers very plausibly grade with the evaluations in mind, and students plausibly evaluate with real or expected grades as a motivator (Eiszler, 2002; Krautmann and Sander, 1999). Impact size and sign has been estimated by using other teacher observables as an instrument in two stage least squares regressions, finding results as predicted by incentive models (Krautmann and Sander, 1999).

Recently, grade inflation has been noted, and there have been efforts to reduce grade inflation. If it is grading higher than other teachers that matters for later outcomes and evaluations, then "inflation" is expected to continue. Schools can implement regulations to attempt to halt or reverse grade inflation, such as grading on a preset curve or with an average grade as a goal (Butcher et al., 2014). Butcher et al. observed the implementation of such a policy at Wellesley, using individual level data. They found results such as the negative impact being more damaging to minorities, and used better controls than most studies such as SAT scores. They found that evaluations were impacted, but not very much.

## 2. Description of the Data

The data is a collection of grades given in courses in the business school at Utah State University, and data collected on the teacher evaluations of the classes. Also included in the dataset is some basic information about the classes, such as instructor and the size of the class. The data begins in 2011, and was collected through 2014. This is the full range of data that was available as of this writing. In 2011, Utah State implemented 'The IDEA System' from the third party 'The IDEA Center' to evaluate teachers. Students complete evaluations at the end of the semester, before grades are received, on the teacher and course. These evaluations are a series of questions with possible responses ranging from 1 to 5, and also give students an opportunity to provide written feedback on the teacher or course. The aggregated teacher evaluations are available to all students within the university, as well as to faculty. The data is used for assessment by the university management of courses and professors, for management and faculty decisions.

Average course GPA is the arithmetic average of the GPA equivalent to the grade each student in the course received. GPA is calculated using the scale used at Utah State University, with 4 being the highest. Only those students who received a final grade are included in the calculation of GPA. This excludes those who withdrew from the class, and those who didn't receive standard grades. This excludes 1.52% of all individual grade records.

### 2.1. *Evaluation Specifics*

Most important to this study is that the students are directly asked in the evaluation to evaluate the quality of the teacher and the course on a 1-5 scale. There are other questions that ask students to evaluate the course on metrics such as: gaining factual knowledge,

learning fundamental principles, learning to apply material, developing teamwork, etc. Additionally, the evaluation includes measures of 'progress on relevant objectives', which are 3 to 5 of the metrics that are chosen by the professor to be 'important', or 'essential', and is a weighted average of these metrics.

The evaluation survey also includes the questions: 'I really wanted to take the course regardless of who taught it?' and 'As a rule, I put forth more effort than other students on academic work.' These questions are used, along with class size, subject, and other data to create 'adjusted scores' of various metrics. An example of this is an overarching summary score that is provided: it utilizes this adjusted data, averaging together other metrics with no clear reason given for the weighting. These adjusted scores are designed to correct for some of the exogenous factors that impact evaluations.

These two questions in particular are interesting because they ask students post hoc to assess their desires and effort. The average 'effort' that students in the data set responded was 3.94, where 3 suggests an average amount of effort. This shows that students do not give unbiased responses.

## *2.2. Grade Policy*

A policy was implemented in the business school at the beginning of 2014 in which there was a suggested ceiling on course GPA. This was a GPA of 3.2 for upper division courses, and 2.8 for lower division courses. Grade data is limited, and does not include identifying information of individuals. This grade ceiling was not strictly enforced, but rather was suggested and encouraged. Some professors were already grading below this standard, and so should not have been impacted by the policy. The number who should have been impacted is shown in table 2.

The policy was communicated directly to professors and not directly to students, nor was it widely publicized. Therefore, we assume that the policy does not impact how students evaluate professors except through the channel of how professors grade. Graphs 1 and 2 demonstrate the distribution of average grades before and after the policy was implemented for the upper division courses. There is no visible shift in the distribution. Graphs 3 and 4 demonstrate the same for the lower division courses. There appears to be a very visible shift. Before the policy, 45.8% of upper division courses and 61.2% of lower division courses were above the later instituted gap, while after the policy shift, 42.5% of upper division courses and 38% of lower division courses were still graded above the cap. The policy seems to have been followed better for lower division courses.

### **3. Empirical Methods**

The main model used in this paper is an instrumental variable approach, where a binary variable representing whether the new grade policy was in effect is used as an instrument for the average course GPA. This method attempts to provide an unbiased estimate of how the grades that teachers give influence the evaluations that students make. As noted in the literature section, teachers may attempt to 'buy' good evaluations by grading leniently, and students may 'reward' teachers who grade leniently with good evaluations. As there is a causality loop, there is a form of endogeneity that may bias results when the impact of grades on evaluations is estimated using a standard least squares model.

As the students complete the evaluations before they actually receive final grades, any correlation between grades and the evaluations will be due to information that students already have of their grades, and their expectations. The amount of information that students have of their grades might vary by course and professor.

There is suspected heterogeneity in the model due to there being many different professors and different courses. This problem is mitigated in the model by including course-professor fixed effects. This creates a fixed effects estimate for courses taught by a single professor. This model thus will not include courses taught by professors who taught only before or after the policy came into place, and will provide fixed effects for courses taught by the same professor both before and after the policy enters effect.

Regressions are also conducted including only lower or upper division courses, or both, due to the differing impact of the policy on the lower and upper divisions.

As follows is how the model would be constructed using OLS:

$$\textit{Evaluation score} = \alpha + \beta \textit{grades} + \textit{controls} + \mu \quad (1)$$

Due to the issue of reverse causation, a two stage least squares model is used, with grades being instrumented by the binary variable of whether the policy was in place. Additionally, courses taught by the same professor that were already beneath the suggested cap are excluded from the regressions, as they should not be impacted by the new policy. They are included separately, for comparison. Results were also obtained for standard OLS and OLS with fixed effects, for means of comparison between the OLS and 2SLS models and also for comparison with other studies.

#### **4. Empirical Results**

The first stage for the upper division of courses, the regression of the post-implementation dummy variable on average course GDP is both insignificant and has a positive sign, yielding a coefficient of 0.021 with a standard error of 0.06. This suggests that the policy had no effect on upper division grading. This could be due to several factors: upper division courses are generally smaller and more personal, and the students are more

likely to be in the major and possibly know the professor personally.

Additionally, the group of teachers that should be unaffected by the policy due to already grading beneath the threshold also do not appear to have been impacted by the policy. The first stage regression also has a positive sign, and is insignificant. (Coefficient: 0.05, standard error: 0.06). This is shown in the 'Robustness' section below.

In table 3, the results of regressions using the teaching quality metric as the dependent variable are presented. They show that the grades are positively correlated with the evaluations, and significantly so. This is consistent with prior research on the subject.

Tables 4 through 7 are the results of two stage least square regressions. The results of two stage least squares regressions of Average Course GPA on the unadjusted Teacher Quality evaluation are in Table 4. Under all of the specifications below, the estimated coefficient is positive and significant at the  $p < .1$  level or less.

Tables 5 through 7 use the same two stage least squares regressions as table 4, but with the left hand side variable being the evaluation of the course as a whole, progress on fundamental indicators, and the summary score indicator respectively. The results are similar to using teacher quality as the independent variable, but with different scale and statistical significance. Notably, the summary score seems to be much more responsive to grades than the other indicators do.

## **5. Robustness Tests**

Table 8 shows the first stage results for professors who should not have been impacted by the policy: those who were already grading below the recommended ceiling. This is a regression using the average course GPA as the independent variable, and the binary variable representing whether the grade policy was in effect.

None of the coefficients have the negative coefficient that would be expected if they had been impacted by the policy, and have coefficients that are statistically insignificant and practically indistinguishable from 0. This suggests that only the group that should have been impacted by the policy was. This also suggests that it is likely that the change in grading that is observed was likely due to the policy, and not some other factor.

## **6. Conclusion and Discussion**

The policy seems to have had an effect on grading, especially in lower division courses. As the policy was not strictly enforced, but was rather more like a guideline, teachers had the discretion to choose to follow the policy or not. Before the policy, 45.8% of upper division courses and 61.2% of lower division course were above the cap, while after the policy was implemented, 42.5% of upper division courses and 38% of lower division courses were still graded above the cap. The first stage results that were obtained suggest that the policy was much more effective for lower division courses.

Using the instrumental variable regression, it appears that the average grades in a course are still positively correlated with the teacher and course evaluations. Coefficient estimates are much larger using the 2SLS approach. The estimates are consistently positive, and with comparable coefficients even with different models, such as not including the teacher-course fixed effects or with weighted regressions.

All of the regressions that use class size weights also yield larger coefficients. This could be due to smaller classes being more personal, or that with higher enrollment teachers are more willing to give lower grades and that evaluations are also negatively correlated with larger class sizes. (NB: I ran regressions, this seems to be the case, and regression results change when 'Enrolled' is included as a control.)



The results of this paper as a whole suggest that there is a positive correlation between grades and teacher evaluations, and using the instrumental variable approach to remove feedback effect bias, suggests that the correlation coefficient is more than twice as large as results obtained through OLS.

While these results are consistent for several different independent variables from the teacher evaluations and with different model specifications, the results may not be externally valid. It is possible that other universities, or even other schools within Utah State University might not have similar properties. Students select schools, and schools select students and have different policies. Also, implicit in the model used in this paper is the assumption that students did not change how they evaluate teachers with the policy change, and how a policy is implemented could cause changes to be path dependent.

However, as far as the results are valid, they suggest that there is a correlation between leniency in grading and how students evaluate the teachers, and that this might be a larger effect than that suggested by an OLS regression. Therefore, the teachers could have an incentive to give artificially high grades to receive better evaluations, which can make a real impact on their careers and also distort management decisions.

The policy implication is that management decisions should possibly take this into account by holding evaluations before grades are given, mandating an institution wide grading standard, or by simply understanding the biases in the evaluation itself. Taking advantage of a policy shift at the business school at Utah State University that introduced suggested caps on the average course grades that teachers gave, an instrumental variable was constructed to predict the average class GPA. Resulting estimated coefficients for the correlation between grades and student teacher evaluations were about twice as large as

obtained from OLS. The correlation between grades and students' evaluations of teachers is positive. This suggests that student evaluations of teachers are biased by the grades that teachers give, which can make them less useful as a gauge of teacher quality. This is consistent with what has been termed Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure." (Strathern, 1997) Management should take possible biases of measures into account.

## References:

IDEA information:

[http://www.usu.edu/aaa/pdf/idea/HO8-InterpretiveGuideDF\\_2Page.pdf](http://www.usu.edu/aaa/pdf/idea/HO8-InterpretiveGuideDF_2Page.pdf)

[http://www.usu.edu/aaa/pdf/idea/Raw\\_vs\\_Adjusted\\_WhichToUse.pdf](http://www.usu.edu/aaa/pdf/idea/Raw_vs_Adjusted_WhichToUse.pdf)

[http://www.usu.edu/aaa/pdf/idea/IDEA\\_FacultyHandout\\_InterpretingResults\\_FINAL.pdf](http://www.usu.edu/aaa/pdf/idea/IDEA_FacultyHandout_InterpretingResults_FINAL.pdf)

GPA:

<http://www.usu.edu/advising/gpa/files/Calculating%20a%20GPA.pdf>

- Braga, Michela, Marco Paccagnella, and Michele Pellizzari. "Evaluating students' evaluations of professors." *Economics of Education Review* 41 (2014): 71-88.
- Butcher, Kristin F., Patrick J. McEwan, and Akila Weerapana. "The effects of an anti-grade-inflation policy at Wellesley College." *Journal of Economic Perspectives* 28.3 (2014): 189-204.
- Centra, John A. "Will teachers receive higher student evaluations by giving higher grades and less course work?." *Research in higher education* 44.5 (2003): 495-518.
- Eiszler, Charles F. "College students' evaluations of teaching and grade inflation." *Research in Higher Education* 43.4 (2002): 483-501.
- Ewing, Andrew M. "Estimating the impact of relative expected grade on student evaluations of teachers." *Economics of Education Review* 31.1 (2012): 141-154.
- Feldman, Kenneth A. "Grades and college students' evaluations of their courses and teachers." *Research in Higher Education* 4.1 (1976): 69-111.
- Holmes, David S. "Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor." *Journal of Educational Psychology* 63.2 (1972): 130.
- Krautmann, Anthony C., and William Sander. "Grades and student evaluations of teachers." *Economics of Education Review* 18.1 (1999): 59-63.
- Kristoff De Witte, and Nicky Rogge. "Accounting for exogenous influences in performance evaluations of teachers." *Economics of Education Review* 30.4 (2011): 641-653.
- Love, David A., and Matthew J. Kotchen. "Grades, course evaluations, and academic incentives." *Eastern Economic Journal* 36.2 (2010): 151-163.
- Marsh, Herbert W., and Lawrence A. Roche. "Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders?." *Journal of educational psychology* 92.1 (2000): 202.

## Appendix: Tables and Figures

Table 1: Selected Summary Statistics

Measure	Mean	Min	Max	Standard Deviation	# of observations
Average Course GPA	3.03	1.88	3.89	0.44	265
Enrolled in Course	73.9	11	269	55.5	262
Teacher Quality	4.2	2.3	5	0.46	261
Course Quality	4.0	2.7	5	0.42	261
effort	3.94	3	4.6	0.22	261

Table 2: Courses falling within grade policy window, by division

	Number of courses teacher pairings graded above ceiling before policy	below	Percent above
Upper and Lower divisions	196	69	74%
Lower Division	122	31	80%
Upper Division	74	38	66%

Table 3: OLS results

	1	2	3	4	5	6
Grade Coefficient (s.e)	.217**(.063)	.203**(.071)	.406**(.102)	.406**(.085)	.305**(.079)	.463**(.095)
Lower division	Both	Lower	Upper	both	both	both
teachcourseid dum				Yes		Yes
Weight by Enrollment					Yes	Yes
Sample size	261	149	112	261	261	261

Table 4: 2SLS Teacher Quality Result

	1	2	3	4	5
First Stage	-.147*(.067)	-.258** (.099)	-.277**(.100 )	-.169**(.043 )	-.147**(.042 )
Coefficient (s.e)	1.024' (.565)	.727'(.394)	1.013'(.521)	.792**(.250)	.799**(.297 )
Lower division	Both	Lower	Lower	Lower	Lower
teachcoursei ddum				Yes	Yes
Weight by Enrollment			Yes		Yes
Sample size	192	118	118	118	118

'p<.10 \*: p<.05 \*\*: p<.01 Robust standard errors

Table 5: 2SLS Course Evaluation Results

CourseRaw

	1	2	3	4	5
First Stage	-.147*(.067)	-.258** (.099)	-.285**(.097 )	-.169**(.043 )	-.177**(.043 )
Coefficient (s.e)	0.78 (.479)	.47(.31)	.738'(.39)	.485'(.25)	.583*(0.255)
Lower division	Both	Lower	Lower	Lower	Lower
teachcoursei ddum				Yes	Yes
Weight by Enrollment			Yes		Yes
Sample size	192	118	118	118	118

Table 6: 2SLS Progress on Fundamental Indicators Evaluation Results

FundRaw

	1	2	3	4	5
First Stage	-.145*(.069)	-.216** (.103)	-.224**(.099)	-.171**(.044)	-.178**(.044)
Coefficient (s.e)	.60 (.42)	.306(.33)	.58(.43)	.33(.28)	.48'(0.27)
Lower division	Both	Lower	Lower	Lower	Lower
teachcourseid dum				Yes	Yes
Weight by Enrollment			Yes		Yes
Sample size	169	106	106	106	106

Table 7: 2SLS Summary Score Evaluation Results

SummaryScore

	1	2	3	4	5
First Stage	-.152*(.066)	-.263** (.098)	-.285**(.096)	-.169**(.043)	-.177**(.043)
Coefficient (s.e)	13.4' (7.4)	8.9'(5.1)	12.9'(7.1)	9.6*(3.8)	10.7*(4.1)
Lower division	Both	Lower	Lower	Lower	Lower
teachcourseid dum				Yes	Yes
Weight by Enrollment			Yes		Yes
Sample size	193	119	119	119	119

Table 8: First stage results

	Both divisions	Lower Division	Upper Division
Coefficient	0.020	0.054	0.158
Robust standard error	0.063	0.058	5.00
P value	0.75	0.36	0.98

Figure 1: Upper division distribution of grades pre-implementation of policy

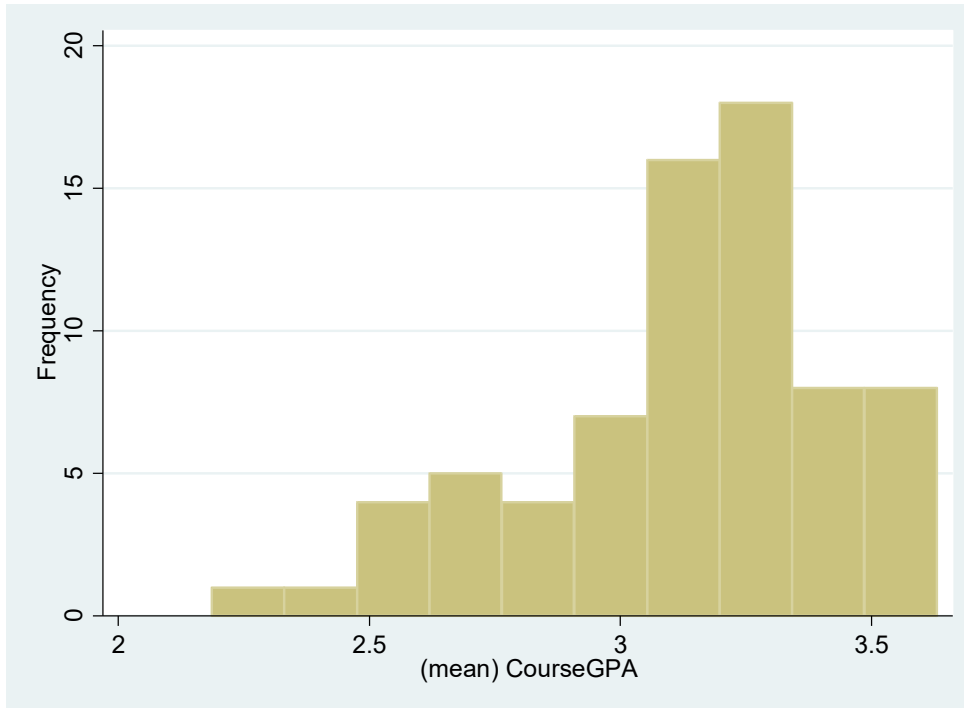


Figure 2: Upper division distribution of grades post-implementation of policy

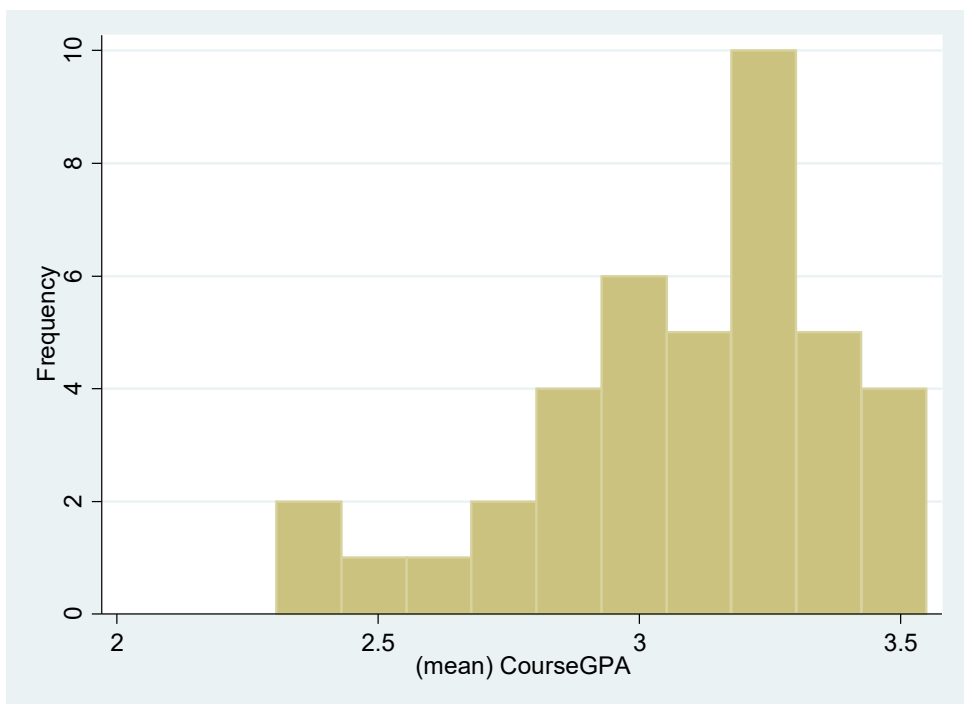




Figure 3: Lower division distribution of grades pre-implementation of policy

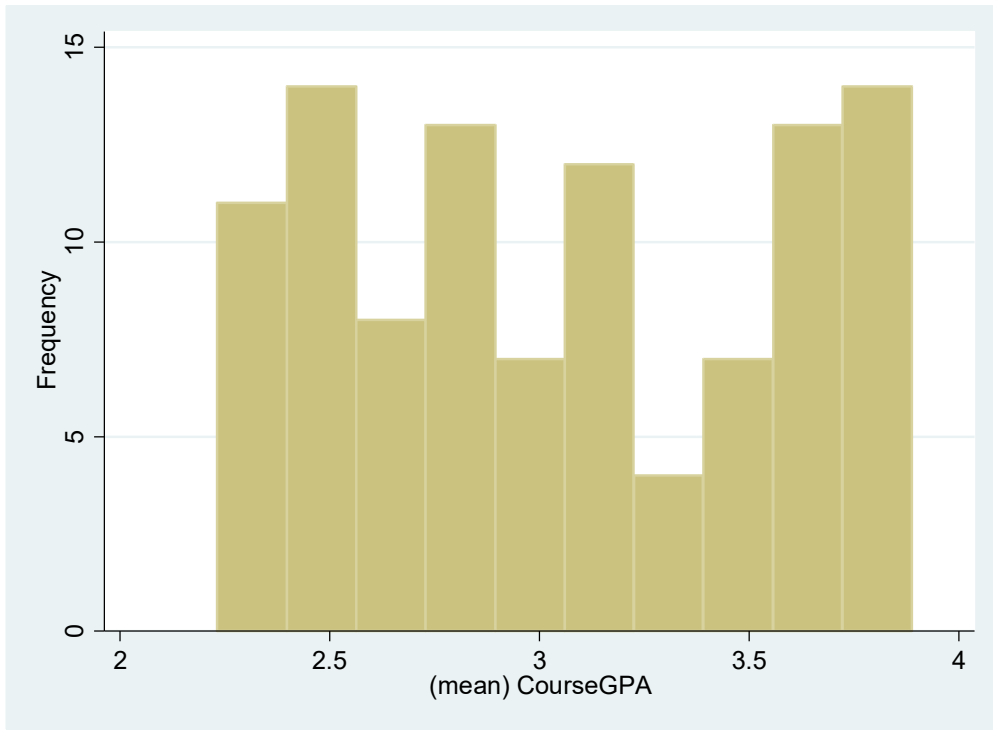


Figure 4: Lower division distribution of grades post-implementation of policy

