

Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching

Zombor Berezvai , Gergely Dániel Lukáts & Roland Molontay

To cite this article: Zombor Berezvai , Gergely Dániel Lukáts & Roland Molontay (2020): Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching, *Assessment & Evaluation in Higher Education*, DOI: [10.1080/02602938.2020.1821866](https://doi.org/10.1080/02602938.2020.1821866)

To link to this article: <https://doi.org/10.1080/02602938.2020.1821866>



Published online: 27 Sep 2020.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching

Zombor Berezvai^{a,b} , Gergely Dániel Lukács^c and Roland Molontay^{c,d}

^aInstitute of Marketing, Corvinus University of Budapest, Budapest, Hungary; ^bInstitute of Economics, Centre for Economic and Regional Studies, Budapest, Hungary; ^cDepartment of Stochastics, Faculty of Natural Sciences, Budapest University of Technology and Economics, Budapest, Hungary; ^dMTA–BME Stochastics Research Group, Budapest, Hungary

ABSTRACT

At many universities, student evaluations of teaching (SET) are used for determining promotion, tenure, and other financial benefits for professors, which gives incentive for them to try to increase their scores. However, previous research, mainly based on US data, indicates that SET scores not only depend on teaching effectiveness and quality, but also on several other factors, most notably on grades. In this study, we contribute to this stream of literature by investigating the effects of grade inflation on SET scores using four years of data from two leading Central European universities. The SET survey is filled out at these universities after the final grades are known. As a methodological novelty, we use weighted regression methods (ordinary least squares, two-stage least squares, fixed effects panel) to account for the differences in class size. The weighting reflects the importance of the average SET score of a class on the evaluation of the instructor. Our findings suggest that increasing the grade of a student by one will cause them to give approximately 0.2–0.4 higher evaluations for the instructor in the SET survey.

KEYWORDS

Higher education; grade inflation; student evaluation of teaching (SET); regression analysis

Introduction

Grade inflation refers to the tendency to lower academic requirements and award higher grades than what students would deserve, based on past requirements. Grade point averages (GPAs) at universities in the USA have been rising steadily for decades. Rojstaczer (2016) examined data up to 2013 on a large sample of public and private US universities and observed an increase of about 0.10–0.15 GPA points per decade, with private universities having a steeper slope. At Harvard University, for example, the percentage of A grades rose from 22% in 1966 to 46% by 1997 (Wilson 1999), which resulted in 91% of the 2001 class graduating with honours (Healy 2001). Although the magnitude of grade inflation appears to be more significant in private, for-profit universities (Baird, Carter, and Roos 2019), it is also observable in public schools. Eiszler (2002) studied two decades of data at a public university in the Midwestern US and found that the percentage of students expecting A grades trended upward noticeably. Universities have been raising the minimum GPAs needed for honours each year, in an effort to combat grade

inflation, but nearly half of students in 2017 still graduated with honours at universities such as Princeton, the University of South Carolina and many others (Korn 2018).

Grade inflation causes grade compression at the upper end of the distribution, which makes it more difficult for top students to stand out. Workplaces may rely on grades to help distinguish skilled students, but due to grade inflation and compression, they do not have sufficient clarity when hiring. This hinders the potential productivity of a company and decreases the incentives for students to study hard in hopes of getting hired to the best companies (Schwager 2012; Nordin, Heckley, and Gerdtham 2019).

The exact causes of grade inflation are highly debated. Winzer (2002) argues that multiple factors simultaneously lead to the increase in grades, such as institutional changes, changes in student demographics (e.g. student entitlement mindset, increase in the ratio of female students and adult learners), the shift of students' majors from low to high grading departments, blurring of the faculty-student relationship, and curriculum changes.

On the other hand, many believe that the growing social and economic relevance of student evaluation of teaching (SET) scores may be one of the leading factors contributing to the grade inflation phenomenon (Isely and Singh 2005; Nowell 2007; Love and Kotchen 2010; Jewell, McPherson, and Tieslau 2013). Student evaluation of teaching rose to prominence in the late twentieth century to monitor the performance of professors. In 20 years, the percentage of universities collecting student evaluations in the USA grew from 29% in 1973 to 86% in 1993 (Seldin 1998). Today, they are used in an overwhelming majority of higher education institutions and many universities rely heavily on SET scores for deciding tenures, promotions and other financial benefits.

Despite its widespread usage, the accuracy of SET scores in determining instructor quality is heavily debated. Academically irrelevant instructor characteristics such as gender, age, and physical attractiveness have shown varying levels of correlation with SET scores in multiple studies (Boring, Ottoboni, and Stark 2016; Boring 2017; Hornstein 2017). Youmans and Jee (2007) observed that giving chocolate to students before filling out the SET questionnaire resulted in substantially higher scores than the control group. Class characteristics, such as the discipline, course credit, and class size can also influence SET scores. Multiple studies have observed that a larger class size negatively affects SET scores (Andrade and Rocha 2012; Ewing 2012; Cho and Cho 2017). In contrast to previous studies, Park and Dooris (2020) showed using a decision tree analysis that the SET instrument has criterion and discriminant validity since course grading policies and the instructor's gender or race/ethnicity have little effect on the SET scores.

On the other hand, some articles even question whether SET scores reflect teaching effectiveness at all (Alauddin and Kifle 2014; Hornstein 2017). Ware and Williams (1975) conducted a famous experiment where an actor, Michael Fox, was coached to give a presentation full of contradictory statements in an enthusiastic manner. The actor received evaluations that were just as high as an actual professor who gave a logical presentation. This phenomenon was later called the 'Dr Fox Effect'. Andrade and Rocha (2012) found that instructors at a Brazilian university who had participated in training programs to improve their quality of teaching did not receive noticeably higher SET scores.

One of the reasons behind the difficulty of assessing and accepting bias in SET scores is the variety of methods and questions used to compile them at different institutions. Most universities use multidimensional SETs, but there is no clear consensus on what exact number and nature of questions provide the highest construct-related validity (Spooren, Brockx, and Mortelmans 2013). Another possible discrepancy is that mandatory SETs encourage rushed, less serious responses from students, while non-mandatory SETs have a lower response rate, and may yield a non-representative pool of student participants (Arthur 2020). The choice of online or paper-based surveys can also affect the response rate of students (Nowell, Gale, and Handley 2010; Spooren, Brockx, and Mortelmans 2013).

However, a factor that is considered to influence SET scores significantly is the grade given to the student by the instructor. At many universities, SETs are filled out before receiving final grades, and so the SET score is influenced by the expected grade. At some other universities, like the ones studied in this paper, the actual grade can influence SET scores as students fill out the surveys after receiving their final grades. It is theorised that some instructors have been giving increasingly higher grades to students in hopes that they receive higher SET scores in exchange.

The fact that at many universities SET scores play an important role in determining promotions, tenures, and other financial benefits gives instructors a clear incentive to try to raise their SET scores. However, instead of improving their teaching effectiveness, professors may be tempted to give better grades or lower academic standards in the hope of less critical evaluations from students. Students need high grades to qualify for scholarships and to get hired for well-paying jobs. Thus, they may reward their instructors with higher SET scores if they received a good grade or take revenge by giving a low SET score if they received a bad grade, even if it accurately reflected their knowledge of the subject. Many believe that this conflict of interest is a leading reason behind the trend of grade inflation and lowering academic standards in many universities (Isely and Singh 2005; Nowell 2007; Love and Kotchen 2010; Jewell, McPherson, and Tieslau 2013).

As a result, the relationship between SET scores and grade inflation has become a significant topic of interest for researchers, with several studies showing strong positive correlation between the SET score given by a student and the grade received from the instructor (Isely and Singh 2005; Nowell 2007; McPherson, Jewell, and Kim 2009; Ewing 2012). Yet, it is difficult to assess whether higher grades are due to grading leniency, or due to actual teaching effectiveness. If SET scores are accurate, then students attending a skilled professor's class would give high SET scores and perform better on tests, even without any grading leniency. Although it is popularly believed that student evaluations contribute to grade inflation, there is no clear consensus on the exact effects, or how accurate SET scores are, in general.

Certain studies have attempted to quantify the bias of SET scores. Ewing (2012) showed on data from the University of Washington that an increase of one grade raises SET scores by around 0.3 to 0.7, depending on the department. Most of the studies on the relationship between SET score and grade inflation have been conducted on data from US universities. However, a few studies have been done elsewhere. In South Korea, Cho and Cho (2017) studied data at the Sungkyunkwan University after a system error resulted in a random selection of students having access to their grades before submitting an instructor evaluation. In Australia, Davies et al. (2007) examined eight years of responses and assessed the quantitative importance of variables outside the instructor's control on SET scores. Bauer and Grave (2011) showed that although German universities have begun to receive indicator-based funding, the reforms have not caused significant grade inflation. The importance of student evaluation has been highlighted in a few Hungarian studies as well: Tóth et al. (2013) identified the satisfaction factors of students from Budapest University of Technology and Economics.

Our goal with this paper is to study the effects of grade distribution on SET scores using four years of data from two leading Hungarian higher education institutions and evaluate the results in comparison to existing literature. The main estimation methods used in this paper include ordinary least squares (OLS) regression, two-stage least squares (TSLS) regression with an instrumental variable chosen to account for the endogeneity of the received grade variable, and fixed effects (FE) panel regression.

Our paper contains three novel elements compared to prior research in this domain. First, as a methodological improvement, we use weighted regression methods to accommodate the differing course sizes and that larger courses are more important in terms of SET scores for the instructors. Second, to the best of our knowledge, this is the first study of this nature from Central Europe, where tertiary education was significantly improved in the past decades and is, therefore, attracting more and more international students (OECD 2019). Third, our data are also

unusual because at the universities in our sample students fill out their SET survey after receiving their grades, while US students generally have to fill them out before.

Data overview

The data used in this study is obtained from the official registers of Budapest University of Technology and Economics (BME) and Corvinus University of Budapest (CUB). For both universities, the data include all faculties.

Grading data are available from the transcripts on a student level. However, student evaluation is anonymous at both universities, therefore, only section level averages are available, and these are used to assess the performance of the instructor. These section level averages were used in this study. For every section in the database, section, course, and instructor characteristics were collected to enrich the database with several control variables.

Since filling out the SET survey is not mandatory, the use of section level averages means that the grades of those students are also considered who did not fill out the SET survey. To mitigate this problem, we filtered out those sections where the response rate of the SET survey was below 40%.

The SET questionnaire contains several questions at both universities. As regards to CUB, we calculated the average of the four questions assessing the instructor's performance during the class. This average is also used at the university for evaluating and rewarding instructors based on their performance. Regarding BME, we used the answers for one question corresponding to the overall performance of the instructor. We calculated an instructor's SET score for a given section by taking the average of all valid responses in that semester. SET scores, similarly to grades, are measured on a 1–5 scale with higher numbers showing better performance.

Certain students may be used to receiving different grades, and while some students may be pleased with a 4, other students might be disappointed. Thus, the grade they receive may influence their evaluations differently, based on their previous academic achievements. The cumulative grade point average (GPA) is a reliable and up-to-date measure of student abilities (Isely and Singh 2005). To take this effect into account, we used the difference between the grade and the cumulative GPA as the variable of interest in our study. The cumulative GPA is calculated including all courses up to the semester in question.

Relative grade shows how much higher the grade from the given course was compared to the previous grades received by the same students. This variable aims to capture the magnitude of grading leniency in a given section. Relative grade can also take larger negative values if the students received much lower grades than their GPA.

Data were collected for eight consecutive semesters at both universities. It ranged from academic year 2013/2014 to academic year 2016/2017 at BME and from academic year 2014/2015 to academic year 2017/2018 at CUB. The difference is caused by the SET system changing at BME in 2017 and at CUB in 2013. The current dataset is therefore free from any systematic changes.

After transforming and combining all the data we received 50,377 semester-course-section-instructor combinations for CUB and 49,065 for BME. However, many observations were deleted due to the very low number of students or very low response rates to SET questionnaires. To summarise, we kept rows that satisfied the following criteria:

- at least 5 valid SET evaluations;
- response rate of the SET questionnaire is at least 40% of the section size, to ensure that the responses are somewhat representative of the entire section;
- the course credit value is between 2 and 9 (i.e. these are regular courses, not for e.g. internships);
- the course requires the instructor to assign a grade at the end of the course, instead of just passing or failing students.

Table 1. Section average statistics for numerical variables.

Variable	University	Mean	Median	Min	Max	SD
SET	BME	4.09	4.21	1.11	5.00	0.62
	CUB	4.21	4.38	1.33	5.00	0.61
Relative grade	BME	0.13	0.10	-2.12	2.71	0.64
	CUB	-0.09	-0.02	-3.90	1.56	0.69
Response rate	BME	0.54	0.52	0.40	1.00	0.11
	CUB	0.70	0.71	0.40	1.00	0.16
Credits	BME	3.80	4.00	2.00	9.00	1.29
	CUB	4.59	5.00	2.00	9.00	1.03

The intersection of these filters resulted in a total of 13,863 semester-course-section-instructor combinations for CUB and 8,646 for BME, which is the final dataset used for the analyses. Most of the dropped observations came from non-full-time training, PhD programs, graded internships and non-core courses (e.g. physical education).

Table 1 shows descriptive statistics on the numerical variables, while Tables 2 and 3 summarise the descriptives on the instructor and section-specific factors. For instructor characteristics, we use gender, age, and rank. Instructor ranks were grouped into five categories: 'Level1', 'Level2', 'Level3', 'Level4', and 'Other'. The 'Other' category mainly contains specific instructors that are outside of the university ranking system such as language teachers, technical instructors and guest lecturers. The levels roughly correspond to PhD students or assistant lecturers, assistant professors, associate professors and university professors.

Methodology

Our main goal is to evaluate the relationship between SET scores and grading and filter out the effect of teaching effectiveness from this relationship. Therefore, we can measure the pure impact of grade on SET scores. As a preliminary step, the plot of SET scores against relative grades is visible in Figure 1. This shows a clear correlation between the two variables.

To control for multiple variables and account for endogeneity, we first test OLS regression, but that may yield biased results, due to unobserved teaching effectiveness. To account for this endogeneity, we also use fixed effects (FE) model, and two-stage least squares (TSLS) regression with a properly chosen instrumental variable.

Each data point in the regressions is a semester-course-section-instructor combination, which is the granularity level available for SET evaluations. As a methodological novelty, these data points in all the regression models are weighted with the number of students in the given section divided by the number of instructors in the given section. Hence, this measures the number of individual level data points used for calculating section level averages. In this way sections with higher number of students will influence the regression results more compared to those sections with few students. This reflects reality as the SET scores for the instructors are calculated by taking all the individual answers into consideration. Moreover, weighted regressions can also mitigate the problem of using section level averages instead of individual student data. Additionally, standard errors are clustered on instructor level in all the regressions to account for the instructors' individual grading habits.

Ordinary least squares

Ordinary least squares (OLS) regression requires the estimation of Equation (1):

$$SET_{i,t} = \beta_0 + \beta_1 \cdot RelGrade_{i,t} + \beta \cdot X_{i,t} + \tau_t + \varepsilon_{i,t} \quad (1)$$

where t denotes different semesters, while i denotes the individual section-instructor combinations within a given semester. β_1 is the coefficient of interest showing the effect of the relative

Table 2. Section specific descriptives.

Factor	Values	Frequency (BME)	Frequency (CUB)			
Section type	Lecture	49%	31%			
	Practice	51%	69%			
Requirement	Mid-year mark	51%	29%			
	Exam mark	49%	71%			
Section level	Undergraduate	73%	63%			
	Graduate	23%	35%			
	Other	5%	2%			
Semester	2013/14/1	17%	–			
	2013/14/2	17%	–			
	2014/15/1	18%	10%			
	2014/15/2	10%	10%			
	2015/16/1	15%	12%			
	2015/16/2	9%	12%			
	2016/17/1	7%	14%			
	2016/17/2	7%	12%			
	2017/18/1	–	16%			
	2017/18/2	–	14%			
Variable	University	Mean	Median	Min	Max	SD
Section Size	BME	53	29	5	708	74
	CUB	45	32	5	818	57

Table 3. Instructor specific descriptives.

Factor		Values		Frequency (BME)	Frequency (CUB)	
Instructor rank	Level1			19%		19%
	Level2			25%		28%
	Level3			31%		27%
	Level4			7%		8%
	Other			18%		19%
Instructor gender	Male			81%		42%
	Female			19%		58%
Variable	University	Mean	Median	Min	Max	SD
Age	BME	46	44	20	88	13
	CUB	43	41	19	87	11

grade (*RelGrade*) on SET scores. $X_{i,t}$ contains all the section- and instructor-specific control variables specified in [Tables 1–3](#). Finally, τ_t and $\varepsilon_{i,t}$ refer to the semester fixed effect and the idiosyncratic error term. For nominal variables, the reference level is always the one mentioned earlier in [Table 2](#) or [3](#).

The semester fixed effects (τ_t) capture the effects of any factors that are fixed within a given semester. This can mean the difference between fall and spring semesters, a new rule applied to the university, or a cohort of newly registered students with different abilities in all classes.

Two-stage least squares

The regression specified in [Equation \(1\)](#) suffers from the problem that an important variable, teaching effectiveness, is left out since it is unobservable. We can assume that this is a positive relationship as better instructors should receive better evaluations. However, teaching effectiveness also influences grade (and thus relative grade); therefore, leaving out teaching effectiveness from the equation can cause an omitted variable bias. The instrumental variable approach provides a potential way to treat this problem.

Ewing (2012) proposed an instrumental variable to be the average grade in all other sections of the same course, not taught by the given instructor. Following this logic, we decided to

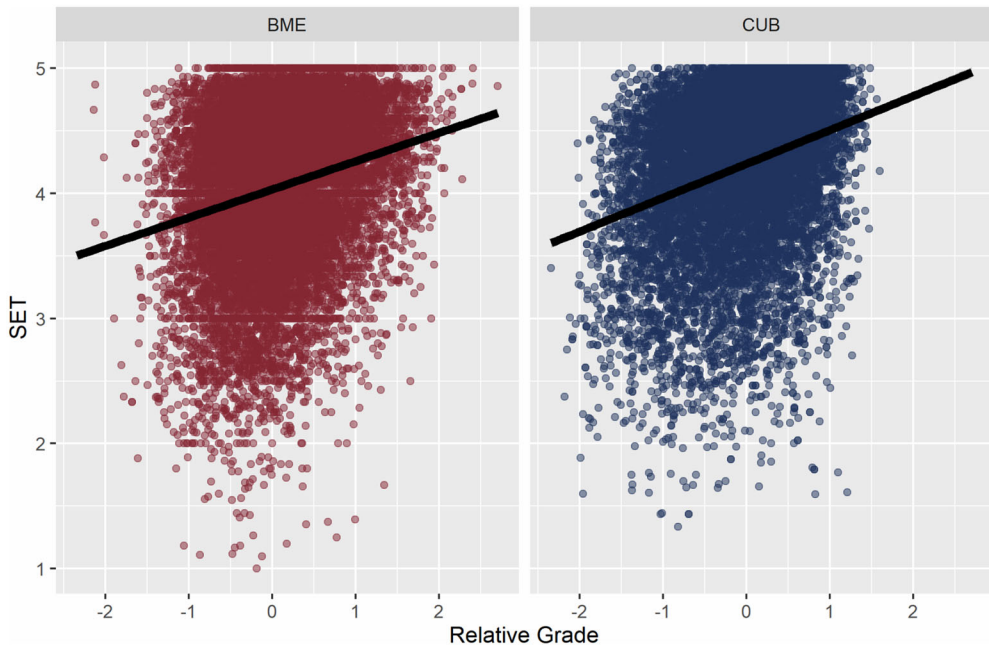


Figure 1. Correlation between relative grade and SET.

choose the instrumental variable to be the average relative grade in all other sections of the same course, with the same section type, where the given instructor is not among the list of instructors.

This instrumental variable is naturally independent of the teaching effectiveness of the given instructor as it contains grades received in other sections not taught by the given instructor. The grades given by other instructors should not tell us anything about the teaching effectiveness of the given instructor, and so we can assume that the correlation is zero. However, the proposed instrumental variable correlates with relative grade mainly due to the difficulty of the underlying course, the similarities in the curriculum, etc.

The endogenous variable and the instrumental variable are indeed correlated, with a Pearson correlation coefficient of 0.82 for BME and 0.86 for CUB. This is further reinforced by all the first stage regressions showing highly significant parameters after controlling for the other independent variables. After an adequate instrumental variable was chosen, two stage least squares method was used for estimation.

The TSLS method can deal with the omitted variable bias; however, the sample size will be significantly lower. In this case, we can only include those sections where there were other sections belonging to the same course taught by other instructors. This reduces the sample size to 3,638 and 6,376 for BME and CUB, respectively.

Fixed effects panel model

Another way to account for unobserved teaching effectiveness is to use a fixed effect (FE) panel model. The FE model is an OLS regression that accounts for all the effects that are constant in time by introducing indicator variables for each instructor. We chose not to include a dummy variable for sections, because we already have multiple section-specific characteristics as control variables.

Using instructor fixed effects, the model picks up the effects of any instructor related characteristics, that are constant in time, on the dependent variable. Because our period spans only

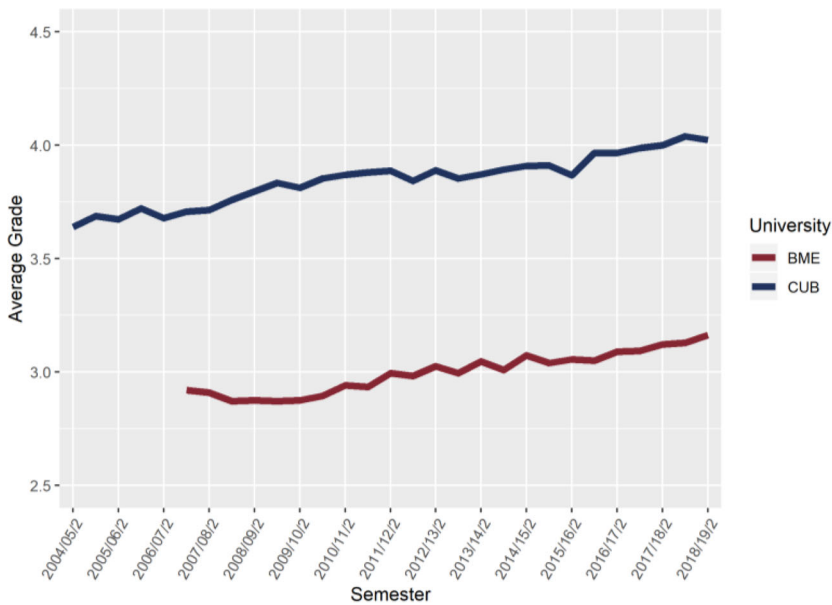


Figure 2. Average grades at the universities analysed.

four years, it is reasonable to assume that general teaching effectiveness is constant for each instructor during the given time period. Due to the short period, the instructor dummies would also capture traits such as physical attractiveness and personality, which could not be captured by the previous methods. However, gender is constant in time, and age does not change significantly in four years, so we removed those variables from the regression equation because they are already captured by the dummy variable for the instructor.

In this regression, β_1 would capture the change in SET scores if a given instructor turns more lenient during the observed four-year time span. Similarly, it would also capture if an instructor introduced more strict requirements or if the difficulty of the examinations varied from previous years. To avoid overfitting, we only kept sections taught by instructors who taught at least one section in at least four semesters during the examined time period. The sample size in this case was 5,109 and 12,204 for BME and CUB, respectively.

Results and discussion

Figure 2 shows the average grades on a longer time frame. Average grades have been rising at both universities, with an approximate increase of 0.3 in the last 10 years.

The average grades for fall semesters are slightly lower, which is likely due to new students mostly joining in September, with poorly performing students dropping out in their first semester.

We examined the grade distributions for different disciplines. The degree of grade inflation varies across disciplines but some levels of grade inflation in the last ten years could be observed in most.

Table 4 contains the estimation results of the OLS regressions. We ran the regression on all data points of both universities and then restricted our samples for the different disciplines (faculties).

The OLS regression estimate for the coefficient of relative grade on the entire dataset is 0.314 for BME and 0.289 for CUB with high significance. This implies that, with all other factors fixed,

Table 4. OLS regression results.

Variables	BME			BME (Economics and Social Sciences)		CUB		CUB (Economics and Business)
	BME (all faculties)	(Natural Sciences)	BME (Engineering)	Social Sciences	CUB (all faculties)	(Social Sciences)		
Relative grade	0.314*** (0.031)	0.470*** (0.072)	0.250*** (0.038)	0.332*** (0.063)	0.289*** (0.036)	0.261*** (0.067)	0.289*** (0.040)	
Response rate	0.031 (0.147)	-0.604* (0.299)	0.276 (0.179)	-0.175 (0.258)	0.153 (0.098)	0.308** (0.153)	0.157 (0.111)	
Course credits	0.038*** (0.014)	0.028 (0.026)	0.047* (0.019)	0.032 (0.028)	0.032* (0.017)	0.004 (0.022)	0.061*** (0.023)	
Exam mark	0.130*** (0.035)	0.152 (0.114)	0.116** (0.042)	0.126* (0.064)	-0.034 (0.045)	-0.025 (0.070)	-0.024 (0.056)	
Practice section	0.079* (0.032)	-0.053 (0.075)	0.117* (0.049)	0.063 (0.055)	-0.102*** (0.031)	-0.134** (0.067)	-0.058* (0.033)	
Level: Graduate	0.061 (0.043)	-0.026 (0.146)	0.032 (0.044)	0.135 (0.075)	0.085** (0.037)	-0.036 (0.076)	0.118*** (0.043)	
Level: Other	0.067 (0.075)	-0.144 (0.133)	0.141 (0.118)	-0.005 (0.098)	0.181*** (0.027)	-0.040 (0.168)	0.227*** (0.060)	
Log(section size)	-0.013 (0.019)	-0.061 (0.035)	-0.005 (0.030)	-0.017 (0.030)	-0.036 (0.027)	-0.085* (0.043)	-0.012 (0.030)	
Instructor rank: Level2	0.002 (0.052)	-0.308* (0.126)	0.087 (0.060)	0.112 (0.096)	0.029 (0.065)	-0.192* (0.104)	0.080 (0.073)	
Instructor rank: Level3	0.033 (0.065)	-0.055 (0.161)	0.031 (0.080)	0.251* (0.116)	0.003 (0.075)	-0.300** (0.122)	0.067 (0.084)	
Instructor rank: Level4	0.210* (0.104)	-0.077 (0.192)	0.291* (0.135)	0.207 (0.146)	-0.091 (0.113)	-0.189 (0.194)	-0.081 (0.128)	
Instructor rank: Other	0.035 (0.063)	-0.069 (0.122)	0.063 (0.080)	0.152 (0.133)	0.030 (0.077)	-0.220** (0.103)	0.055 (0.103)	
Female instructor	-0.040 (0.057)	0.105 (0.100)	-0.053 (0.111)	-0.180* (0.089)	0.016 (0.054)	-0.080 (0.096)	0.028 (0.062)	
Instructor age	-0.008 (0.013)	0.058* (0.027)	-0.021 (0.016)	-0.045 (0.024)	0.012 (0.018)	0.025 (0.022)	0.008 (0.023)	
Instructor age squared	-0.000 (0.000)	-0.001** (0.000)	0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	
R ²	0.156	0.222	0.158	0.251	0.161	0.172	0.171	
Num. obs.	8,646	1,633	5,714	1,299	13,863	3,389	10,474	

Notes: All regressions contained an intercept and semester fixed effects. Cluster robust standard errors are in parentheses. Significance levels: ****p* < 0.01; ***p* < 0.05; **p* < 0.1.

Notes: All regressions contained an intercept and semester fixed effects. Cluster robust standard errors are in parentheses. Significance levels: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

an instructor lowering requirements and giving their students one grade higher would increase their SET scores by roughly 0.3 on average. It is important to note that this coefficient might be biased due to omitting teaching effectiveness as explained earlier.

When looking at the disciplines separately, the coefficient of relative grade is the highest for natural sciences (0.470), but the estimate for all other disciplines ranges from 0.25 to 0.35. The lowest is for engineering. The Faculty of Natural Sciences teaches several compulsory calculus courses to thousands of engineering students. These courses have a reputation for having strict requirements and failing a significant number of students each year. The average grade for courses taught by the faculty is significantly lower; thus, it is understandable that students in these courses have a higher appreciation for grading leniency.

CUB data show that the effect of grade on SET scores is similar in the Faculties of Economics and Business, and Social Sciences. CUB results are the closest to the results of the Faculty of Economics and Social Sciences of BME. This shows that students studying the same disciplines behave similarly, strengthening our results and improving their general applicability.

Table 5 shows the estimation results for the TSLS regressions. After adjusting for the endogeneity of the relative grade variable, the TSLS regression estimate shows a coefficient of 0.291 for BME with high significance, which is close to the estimate of the OLS regression. According to the Wu–Hausman endogeneity test, OLS is not biased, so it can be used as an efficient estimator in the case of BME. Due to the lack of more than one section in several courses, absolute sample sizes are much lower than for OLS. This is especially problematic for the Faculties of Natural Sciences, and Economics and Social Sciences. TSLS coefficients are rather high in these cases; however, these might not be representative for the entire faculty as the sample size is lower by 59% and 78% compared to OLS, respectively.

The estimated coefficient of CUB (0.211) shows a somewhat lower relationship; nevertheless, it is still highly significant. Results by faculty are very similar to the overall results, however, it is only significant for the Faculty of Economics and Business. This might be due to the relatively low sample size (only around 1,000 and 66% lower compared to OLS) of the Faculty of Social Sciences. Endogeneity test results indicate a significant difference between OLS and TSLS results, i.e. omitted variable bias is confirmed for CUB. This suggests that leaving teaching effectiveness out of the analysis is biasing results for some universities, but not for all.

The results of the fixed effects (FE) regressions can be seen in Table 6. Sample sizes are closer to the OLS ones in this case. The coefficient of relative grade estimated by the FE regression is slightly less than the OLS one, but still above 0.2 for both universities. Similarly to the previous regressions, we can also observe here that the estimate for the Faculty of Natural Sciences at BME (0.362) is the highest, but the estimates for all the disciplines are highly significant and close to each other. Interestingly, CUB results indicate identical impact of grade on SET scores (0.228) as BME results (0.229). Additionally, results are close to the TSLS results indicating that using instructor fixed effects we can also control for unobserved teaching effectiveness; i.e. it is constant in time once a shorter time frame is considered.

Using any of the three regression methods, the effect of relative grade on SET scores for both universities is estimated to be around 0.2–0.4. OLS results seemed to be biased in some cases. TSLS and FE results show a generally lower effect; however, the difference is not major from a policy perspective. This is in line with most of the results observed in other studies.

An unusual feature of our study is that at the universities in our sample students fill out evaluations after receiving their final grades. At other universities student evaluations are filled out before receiving a grade. Therefore, the majority of studies has used expected grade. Ewing (2012) estimated the coefficient for the relative expected grade to be 0.237 using fixed effects, and 0.325 using OLS regression. McPherson, Jewell, and Kim (2009) used a random effects model and arrived at a coefficient of 0.271 for their expected grade variable on principles courses, using data from the University of North Texas. Langbein (2008) estimated a 0.57 expected grade

Table 5. TSLS regression results.

Variables	BME			BME (Economics and Social Sciences)		CUB		CUB (Economics and Business)
	BME (all faculties)	(Natural Sciences)	BME (Engineering)	Social Sciences	CUB (all faculties)	(Social Sciences)		
Relative grade	0.291*** (0.084)	0.569** (0.251)	0.186* (0.102)	0.420*** (0.092)	0.211*** (0.049)	0.193 (0.139)		0.199*** (0.051)
Response rate	0.084 (0.199)	-0.461 (0.381)	0.291 (0.212)	-0.330 (0.310)	0.255** (0.118)	0.241 (0.164)		0.297** (0.130)
Course credits	0.071*** (0.020)	0.042 (0.039)	0.086*** (0.024)	0.129* (0.068)	0.055** (0.024)	0.009 (0.033)		0.075** (0.032)
Exam mark	0.158** (0.070)	0.196 (0.132)	0.130* (0.071)	0.212** (0.083)	-0.013 (0.052)	0.015 (0.135)		0.029 (0.059)
Practice section	0.155*** (0.068)	0.011 (0.164)	0.304*** (0.109)	0.042 (0.112)	-0.297*** (0.100)	-0.153 (0.206)		-0.235** (0.102)
Level: Graduate	0.001 (0.107)	0.003 (0.172)	0.051 (0.081)	0.558*** (0.183)	0.038 (0.045)	-0.200* (0.103)		0.068 (0.049)
Level: Other	0.188** (0.075)	-	0.178** (0.075)	0.072 (0.112)	0.272* (0.047)	-		0.319** (0.149)
Log(section size)	-0.034 (0.035)	-0.093 (0.071)	0.047 (0.052)	-0.145*** (0.051)	-0.125*** (0.047)	-0.157* (0.089)		-0.090* (0.047)
Instructor rank: Level2	0.031 (0.074)	-0.351** (0.166)	0.068 (0.073)	-0.039 (0.148)	-0.061 (0.070)	-0.286*** (0.105)		-0.025 (0.075)
Instructor rank: Level3	0.034 (0.101)	0.010 (0.177)	-0.009 (0.106)	0.628*** (0.152)	-0.025 (0.084)	-0.318** (0.138)		0.019 (0.093)
Instructor rank: Level4	0.483* (0.250)	0.523*** (0.182)	0.452* (0.247)	0.266 (0.167)	-0.207* (0.107)	-0.530*** (0.214)		-0.139 (0.115)
Instructor rank: Other	-0.087 (0.074)	-0.141 (0.135)	-0.076 (0.082)	-0.132 (0.111)	-0.017 (0.069)	-0.276*** (0.066)		-0.043 (0.090)
Female instructor	0.040 (0.074)	0.150 (0.118)	-0.007 (0.110)	-0.379** (0.147)	0.061 (0.061)	-0.064 (0.073)		0.059 (0.068)
Instructor age	-0.027 (0.020)	0.016 (0.030)	-0.034 (0.022)	-0.005 (0.032)	0.034* (0.020)	0.036 (0.024)		0.028 (0.025)
Instructor age squared	0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)	-0.0004** (0.0002)	-0.000 (0.000)		-0.000 (0.000)
R ²	0.198	0.274	0.179	0.593	0.175	0.190		0.171
First stage F-test	5,092.23	257.03	4,119.21	106.04	1,389.42	97.71		1,239.43
Num. obs.	3,638	664	2,691	283	6,376	1,147		5,229

Notes: All regressions contained an intercept and semester fixed effects. Cluster robust standard errors are in parentheses. Significance levels: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table 6. Fixed effect regression results.

Variables	BME (Economics and Social Sciences)				CUB (Social Sciences)			CUB (Economics and Business)
	BME (all faculties)	BME (Natural Sciences)	BME (Engineering)	BME (Economics and Social Sciences)	CUB (all faculties)	CUB (Social Sciences)	CUB (Economics and Business)	
Relative grade	0.229*** (0.026)	0.362*** (0.090)	0.212*** (0.026)	0.197*** (0.058)	0.228*** (0.018)	0.213*** (0.029)	0.234*** (0.022)	
Response rate	0.023 (0.083)	0.165 (0.169)	0.055 (0.105)	-0.359 (0.196)	-0.155*** (0.049)	-0.077 (0.090)	-0.176*** (0.056)	
Course credits	0.015 (0.013)	0.070* (0.027)	0.005 (0.016)	-0.003 (0.022)	0.013 (0.008)	0.014 (0.011)	0.012 (0.011)	
Exam mark	0.038 (0.030)	0.071 (0.099)	0.044 (0.035)	-0.037 (0.065)	0.050** (0.023)	0.049 (0.039)	0.045 (0.028)	
Practice section	0.028 (0.022)	0.096 (0.056)	-0.008 (0.032)	0.036 (0.037)	-0.064*** (0.013)	-0.110*** (0.025)	-0.041*** (0.015)	
Level: Graduate	0.030 (0.033)	0.069 (0.112)	0.039 (0.043)	0.009 (0.050)	0.057** (0.024)	-0.011 (0.042)	0.082*** (0.029)	
Level: Other	0.072 (0.056)	0.435*** (0.095)	0.011 (0.095)	0.110 (0.068)	0.154*** (0.057)	-0.249 (0.158)	0.203*** (0.057)	
Log(section size)	-0.093*** (0.015)	-0.075 (0.042)	-0.110*** (0.020)	-0.078*** (0.022)	-0.091*** (0.011)	-0.118*** (0.021)	-0.080*** (0.013)	
Instructor rank: Level2	0.054 (0.055)	-0.172 (0.205)	0.109 (0.092)	0.049 (0.068)	0.025 (0.040)	-0.022 (0.059)	0.027 (0.043)	
Instructor rank: Level3	0.107 (0.079)	-0.068 (0.230)	0.194 (0.124)	-0.036 (0.109)	0.005 (0.061)	0.005 (0.087)	0.003 (0.071)	
Instructor rank: Level4	0.138 (0.093)	-0.166 (0.226)	0.265* (0.135)	-	-0.059 (0.085)	-0.231* (0.126)	-0.0002 (0.081)	
Instructor rank: Other	-0.136 (0.087)	-0.318 (0.220)	-0.187 (0.137)	0.032 (0.097)	-0.118* (0.063)	-0.161** (0.066)	-0.116 (0.072)	
R ²	0.802	0.797	0.814	0.780	0.732	0.735	0.735	
Num. obs.	5,109	876	3,414	819	12,204	2,987	9,217	

Notes: All regressions contained semester and instructor fixed effects. Cluster robust standard errors are in parentheses. Significance levels: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

coefficient using fixed effects method; however, their SET system uses a 6-point scale, which is part of the reason why their coefficient is larger.

The fact that we see little difference in the magnitude of the effect of grade on SET scores in our sample compared to related studies despite the difference in the timing of the teaching evaluation suggests that timing has little effect since students can efficiently predict their grades. This could be due to coursework gaining popularity at universities. Another possible explanation could be that US students are more obsessed with their grades and the higher appreciation for grading leniency leads to a high effect even if they are not aware of their actual grade (Thomas 2011).

It is also worth observing the importance of the control variables. Response rate seems to be insignificant in all models for BME, but it is quite significant for CUB in the TSLS and FE models. However, the response rate is much higher on average at CUB (Table 1).

Section size seems to have a very significant, increasingly negative impact on the SET score at both universities based on the TSLS and FE models. This suggests that after controlling for teaching effectiveness, larger section size decreases the quality of teaching. This is in line with results from multiple studies, such as Langbein (2008), Ewing (2012) and Cho and Cho (2017), who also found that the class size and enrolment rate have a negative impact on evaluation scores. This result is as expected because larger classes are more difficult to teach and instructors with larger sections cannot establish a personal connection with each student which would likely result in better evaluations.

The credit value of the course is positively impacting SET scores; however, the results are often not significant. This could be because students put more effort into more important courses; hence, they pay more attention to the instructor that could enhance their evaluation of the instructor. However, practice sections, where the number of students is normally lower and instructors can better provide individual support for students, are having a negative impact on SET scores at CUB. This could be connected to enhanced expectations towards the instructors in terms of individual support and help. On the contrary, at the engineering faculties of BME, the effect of the practice section is positive. One reason could be that engineering students are more practical in general and, therefore, they may prefer practical courses over theoretical ones.

Surprisingly, gender, age, and university rank are not significant in the majority of the regressions indicating that demographic factors are not affecting SET scores in Hungary.

Conclusion

We studied the effect of grades on the evaluation of the instructor given by the students in the SET survey in case of two leading Hungarian universities. Our paper is the first that applied Central European data to analyse this phenomenon. This region is growing in importance in global higher education with more and more local and foreign students starting university studies (OECD 2019). A specialty in the dataset is that SET surveys are filled out at both universities after the grades are known.

We used three different estimation methods (OLS, TSLS and FE) with cluster robust standard errors and weighting with section size to get robust and reliable estimates.

As a limitation of the study, due to the anonymity of the SET surveys, we had to use section level average data instead of individual data. This can cause a problem if highly satisfied or dissatisfied students fill out the SET surveys more likely than others. However, to minimise the impact of this on the results, sections with low SET response rates were eliminated as the bias can be more present in those cases. Furthermore, the weighted regression methodology can also help to overcome this problem as it places more importance on large sections. Finally, several other studies (e.g. Langbein 2008; Ewing 2012) aggregated the individual data on section

level, therefore, our approach is in line with them. Therefore, we believe that it is not biasing our results materially.

From the results of each of the three regression models, we can conclude that grading leniency does indeed have a significant impact on SET scores at both universities. Our results showed that, on average, increasing the grade of a student by one will cause them to give approximately 0.2–0.4 higher evaluations on the SET survey for the given instructor. This means that instructors are able to buy better evaluations with lenient grading. However, it does not imply that instructors purposely apply grading leniency to increase their SET scores, it only estimates how easy it would be to do so if they chose to attempt it.

Results are similar in magnitude to the findings of previous studies despite the majority of these studies using expected grade as the final grade was not known by the students at the time they filled out the SET survey. However, in the US, coursework constitutes an important part of the evaluation, therefore, students know several elements of their grades already. This can constitute the basis for their estimate of the expected grade. Our findings, therefore, reinforce that the bias in SET scores is present regardless of the timing of the survey.

Our results also suggest that the SET results in Central Europe could be biased similarly to evaluation results around the world. There can be several ways to overcome this bias. Other articles have suggested various methods involving adjusting the SET scores with a function of the grade average in the class (Ewing 2012), or student GPA (Koper et al. 2015), and controlling for class and faculty characteristics (Zangenehzadeh 1988). Stark and Freishtat (2014) highlighted in their review that more emphasis should be placed on teaching portfolios and classroom observations instead of SET scores.

A potential way to obtain an unbiased SET score would be to subtract 0.3 or 0.4 multiplied by the difference in course grade average compared to the average grade of all students in the entire faculty. This would discourage instructors from lenient grading in the hopes of obtaining higher SET scores. We also believe that an important potential change to the SET survey would be to include a question that asks students directly to evaluate the magnitude of grading leniency that they experienced, compared to other courses. This would be the first step towards achieving unbiased SET results and allow the possibility of more direct studies in the future.

Another possible solution can be to put more emphasis on teaching portfolios, classroom observation and peer-review of teaching suggested by Stark and Freishtat (2014). This is in line with the practice of the Faculty of Economics and Social Sciences at BME, which has introduced the use of faculty evaluations of teaching as a more reliable measure of teaching effectiveness (Andor, Tóth, and Árva 2018). These evaluations can reassure instructors about their strengths and help them recognise areas of improvement while providing each other with useful advice. Moreover, a comprehensive university peer review of teaching programs could also provide balanced feedback to the academic staff and management on best practices and common issues. Harrison et al. (2020) in their recent meta-review provided an overview on the opportunities and challenges relating to the most common quality evaluation and enhancement methods of teaching, and they found that a multi-modal approach may be the most effective. Therefore, an important recommendation is that – despite it being a widespread practice all over the world – universities should not solely rely on student evaluation of teaching scores for assessing the quality of their professors' teaching practice.

Acknowledgements

We would like to thank Mihály Szabó, Bálint Csabay, Gergely Kováts and Imre Sebestyén for their help in data collection and interpretation. We are also grateful for Dániel Horn and Hedvig Horváth for their comments on previous versions of this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research of Zombor Berezvai was supported by the ÚNKP-20-4 New National Excellence Program of the Ministry of Innovation and Technology. The research of Gergely Dániel Lukáts was supported by the ÚNKP-19-2 New National Excellence Program of the Ministry of Innovation and Technology. The research of Roland Molontay was supported by the ÚNKP-19-3 New National Excellence Program of the Ministry of Innovation and Technology; and the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Artificial Intelligence research area of Budapest University of Technology and Economics (BME FIKP-MI/SC).

Notes on contributors

Zombor Berezvai is assistant lecturer at Corvinus University of Budapest and the Chief Economist at the Hungarian Competition Authority. His main research interest includes retail innovation, empirical modelling related to industrial organization, and quality development in higher education. He conducted several analyses to the Directorate of Strategy and Quality Development of Corvinus University of Budapest to improve the motivation system of the instructors.

Gergely Dániel Lukáts earned his MSc degree in applied mathematics from the Budapest University of Technology and Economics. He was awarded with prestigious prizes during his studies, including a 1st prize at the National Scientific Students' Associations Conference. Currently he works as an internal auditor at Morgan Stanley.

Roland Molontay is a junior research fellow at the MTA–BME Stochastics Research Group operating at the Department of Stochastics at Budapest University of Technology and Economics. He earned his MSc degree in applied mathematics with highest honors. His PhD thesis focuses on network theory. He is also interested in data science and applied probability. He leads a small research group conducting data-driven educational research.

ORCID

Zombor Berezvai  <http://orcid.org/0000-0001-7807-2977>

References

- Alauddin, M., and T. Kifle. 2014. "Does the Student Evaluation of Teaching Instrument Really Measure Instructors' Teaching Effectiveness? An Econometric Analysis of Students' Perceptions in Economics Courses." *Economic Analysis and Policy* 44 (2): 156–168. doi:[10.1016/j.eap.2014.05.009](https://doi.org/10.1016/j.eap.2014.05.009).
- Andor, Gy., Zs E. Tóth, and G. Árvai. 2018. "Critical to Quality Aspects of University Teaching from the Students' Point of View – Experience of University Peer Review of Teaching Program." In *EDULEARN18 Proceedings*, edited by L. Gómez Chova, A. López Martínez, and I. Candel Torres, 4344–4352. Palma: IATED Academy. doi:[10.21125/edulearn.2018.1092](https://doi.org/10.21125/edulearn.2018.1092).
- Andrade, E. C., and B. P. Rocha. 2012. "Factors Affecting the Student Evaluation of Teaching Scores: Evidence from Panel Data Estimation." *Estudos Econômicos (São Paulo)* 42 (1): 129–150. doi:[10.1590/S0101-41612012000100005](https://doi.org/10.1590/S0101-41612012000100005).
- Arthur, L. 2020. "Evaluating Student Satisfaction – Restricting Lecturer Professionalism: Outcomes of Using the UK National Student Survey Questionnaire for Internal Student Evaluation of Teaching." *Assessment & Evaluation in Higher Education* 45 (3): 331–344. doi:[10.1080/02602938.2019.1640863](https://doi.org/10.1080/02602938.2019.1640863).
- Baird, A. F., J. S. Carter, and J. M. Roos. 2019. "Seeking Evidence of Grade Inflation at For-Profit Colleges and Universities." *Sociological Focus* 52 (4): 343–358. doi:[10.1080/00380237.2019.1668321](https://doi.org/10.1080/00380237.2019.1668321).
- Bauer, T. K., and B. S. Grave. 2011. *Performance-Related Funding of Universities – Does More Competition Lead to Grade Inflation?* Ruhr Economic Papers No. 288. Essen: Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI). doi:[10.2139/ssrn.1954693](https://doi.org/10.2139/ssrn.1954693).
- Boring, A. 2017. "Gender Biases in Student Evaluations of Teaching." *Journal of Public Economics* 145 (1): 27–41. doi:[10.1016/j.jpubeco.2016.11.006](https://doi.org/10.1016/j.jpubeco.2016.11.006).
- Boring, A., K. Ottoboni, and P. B. Stark. 2016. "Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness." *ScienceOpen Research* 2016: 1–11.

- Cho, D., and J. Cho. 2017. "Does More Accurate Knowledge of Course Grade Impact Teaching Evaluation?" *Education Finance and Policy* 12 (2): 224–240. doi:10.1162/EDFP_a_00197.
- Davies, M., J. Hirschberg, J. Lye, C. Johnston, and I. McDonald. 2007. "Systematic Influences on Teaching Evaluations: The Case for Caution." *Australian Economic Papers* 46 (1): 18–38. doi:10.1111/j.1467-8454.2007.00303.x.
- Eiszler, C. F. 2002. "College Students' Evaluations of Teaching and Grade Inflation." *Research in Higher Education* 43 (4): 483–501. doi:10.1023/A:1015579817194.
- Ewing, A. M. 2012. "Estimating the Impact of Relative Expected Grade on Student Evaluations of Teachers." *Economics of Education Review* 31 (1): 141–154. doi:10.1016/j.econedurev.2011.10.002.
- Harrison, R., L. Meyer, P. Rawstorne, H. Razee, U. Chitkara, S. Mears, and C. Balasooriya. 2020. "Evaluating and Enhancing Quality in Higher Education Teaching Practice: A Meta-Review." *Studies in Higher Education*. Advance online publication. doi:10.1080/03075079.2020.1730315.
- Healy, P. 2001. "Harvard's Honors Fall to the Merely Average." *Boston Globe*, October 8, A1.
- Hornstein, H. A. 2017. "Student Evaluations of Teaching Are an Inadequate Assessment Tool for Evaluating Faculty Performance." *Cogent Education* 4 (1): 1–8. doi:10.1080/2331186X.2017.1304016.
- Isely, P., and H. Singh. 2005. "Do Higher Grades Lead to Favorable Student Evaluations?" *The Journal of Economic Education* 36 (1): 29–42. doi:10.3200/JECE.36.1.29-42.
- Jewell, R. T., M. A. McPherson, and M. A. Tieslau. 2013. "Whose Fault Is It? Assigning Blame for Grade Inflation in Higher Education." *Applied Economics* 45 (9): 1185–1200. doi:10.1080/00036846.2011.621884.
- Koper, P. T., J. Felton, K. J. Sanney, and J. B. Mitchell. 2015. "Real GPA and Real SET: Two Antidotes to Greed, Sloth and Cowardice in the College Classroom." *Assessment & Evaluation in Higher Education* 40 (2): 248–264. doi:10.1080/02602938.2014.904272.
- Korn, M. 2018. "You Graduated Cum Laude? So Did Everyone Else." *The Wall Street Journal*, July 2. <https://www.wsj.com/articles/you-graduated-cum-laude-so-did-everyone-else-1530523801>.
- Langbein, L. 2008. "Management by Results: Student Evaluation of Faculty Teaching and the Mis-Measurement of Performance." *Economics of Education Review* 27 (4): 417–428. doi:10.1016/j.econedurev.2006.12.003.
- Love, D. A., and M. J. Kotchen. 2010. "Grades, Course Evaluations, and Academic Incentives." *Eastern Economic Journal* 36 (2): 151–163. doi:10.1057/eej.2009.6.
- McPherson, M. A., R. T. Jewell, and M. Kim. 2009. "What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes." *Eastern Economic Journal* 35 (1): 37–51. doi:10.1057/palgrave.eej.9050042.
- Nordin, M., G. Heckley, and U. Gerdtham. 2019. "The Impact of Grade Inflation on Higher Education Enrolment and Earnings." *Economics of Education Review* 73 (101936): 101936–101910. doi:10.1016/j.econedurev.2019.101936.
- Nowell, C. 2007. "The Impact of Relative Grade Expectations on Student Evaluation of Teaching." *International Review of Economics Education* 6 (2): 42–56. doi:10.1016/S1477-3880(15)30104-3.
- Nowell, C., L. R. Gale, and B. Handley. 2010. "Assessing Faculty Performance Using Student Evaluations of Teaching in an Uncontrolled Setting." *Assessment & Evaluation in Higher Education* 35 (4): 463–475. doi:10.1080/02602930902862875.
- OECD. 2019. *Education at a Glance 2019: OECD Indicators*. Paris: OECD Publishing. doi:10.1787/f8d7880d-en.
- Park, E., and J. Dooris. 2020. "Predicting Student Evaluations of Teaching Using Decision Tree Analysis." *Assessment & Evaluation in Higher Education* 45 (5): 776–793. doi:10.1080/02602938.2019.1697798.
- Rojstaczer, S. 2016. "Grade Inflation at American Colleges and Universities." <http://www.gradeinflation.com/>.
- Schwager, R. 2012. "Grade Inflation, Social Background, and Labour Market Matching." *Journal of Economic Behavior & Organization* 82 (1): 56–66. doi:10.1016/j.jebo.2011.12.012.
- Seldin, P. 1998. "How Colleges Evaluate Teaching: 1988 vs. 1998: Practices and Trends in the Evaluation of Faculty Performance." *AAHE Bulletin* 50 (7): 3–7.
- Spooren, P., B. Brockx, and D. Mortelmans. 2013. "On the Validity of Student Evaluation of Teaching: The State of the Art." *Review of Educational Research* 83 (4): 598–642. doi:10.3102/0034654313496870.
- Stark, P. B., and R. Freishtat. 2014. "An Evaluation of Course Evaluations." *ScienceOpen Research* 2014: 1–7.
- Thomas, J. 2011. "Factors Relating to Student Grade Obsession: A Quantitative Correlational Study." PhD diss., University of Phoenix.
- Tóth, Zs. E., T. Jónás, R. Bérces, and B. Bedzsula. 2013. "Course Evaluation by Importance-Performance Analysis and Improving Actions at the Budapest University of Technology and Economics." *International Journal of Quality and Service Sciences* 5 (1): 66–85. doi:10.1108/17566691311316257.
- Ware, J. E., and R. G. Williams. 1975. "The Dr. Fox Effect: A Study of Lecturer Effectiveness and Ratings of Instruction." *Journal of Medical Education* 50 (2): 149–156. doi:10.1097/00001888-197502000-00006.
- Wilson, B. P. 1999. "The Phenomenon of Grade Inflation in Higher Education." *Phi Kappa Phi Forum* 79 (4): 38–41.
- Winzer, M. 2002. "Grade Inflation: An Appraisal of the Research." *Faculty of Education University of Lethbridge*. <https://essentialedits.ca/inflation/short.htm>.
- Youmans, R. J., and B. D. Jee. 2007. "Fudging the Numbers: Distributing Chocolate Influences Student Evaluations of an Undergraduate Course." *Teaching of Psychology* 34 (4): 245–247. doi:10.1080/00986280701700318.
- Zangenehzadeh, H. 1988. "Grade Inflation: A Way Out." *Journal of Economic Education* 19 (3): 217–226. doi:10.1080/00220485.1988.10845263.