# Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation

Stroebe, Wolfgang

*Published in:*
Basic and Applied Social Psychology

*DOI:*
10.1080/01973533.2020.1756817

Basic and Applied Social Psychology

ISSN: 0197-3533 (Print) 1532-4834 (Online) Journal homepage: https://www.tandfonline.com/loi/hbas20

# Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis

Wolfgang Stroebe

**To cite this article:** Wolfgang Stroebe (2020) Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis, Basic and Applied Social Psychology, 42:4, 276-294, DOI: 10.1080/01973533.2020.1756817

**To link to this article:** https://doi.org/10.1080/01973533.2020.1756817

BASIC AND APPLIED SOCIAL PSYCHOLOGY
2020, VOL. 42, NO. 4, 276–294
https://doi.org/10.1080/01973533.2020.1756817



# Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis

Wolfgang Stroebe

University of Groningen

**ABSTRACT**

Student Evaluations of Teaching (SETs) do not measure teaching effectiveness, and their widespread use by university administrators in decisions about faculty hiring, promotions, and merit increases encourages poor teaching and causes grade inflation. Students need to get good grades, and faculty members need to get good SETs. Therefore, SETs empower students to shape faculty behavior. This power can be used to reward lenient-grading instructors who require little work and to punish strict-grading instructors. This article reviews research that shows that students (a) reward teachers who grade leniently with positive SETs, (b) reward easy courses with positive SETs, and (c) choose courses that promise good grades. The study also shows that instructors want (and need) good SETs.

Student Evaluations of Teaching (SETs) were independently developed in the 1920s by the educational psychologist Herman H. Remmers at Purdue University (e.g., Remmers & Brandenburg, 1927) and the learning psychologist Edwin R. Guthrie (e.g., Guthrie, 1953) at the University of Washington. Remmers and Guthrie wanted to provide university teachers with information about how their teaching was perceived by students and thus help them to make improvements, where necessary. They intended to limit access to these course evaluations to course teachers. Even though Guthrie warned in 1953 that "it would be a serious misuse of this information to accept it as ultimate measure of merit" (p. 221), SETs soon became valued sources of information for university administrators, who used them as a basis for decisions about merit increases and promotion.

The typical SET consists of forms that ask students to rate their perception of course teachers, often on 5-point Likert scales, ranging from *strongly agree* to *strongly disagree*. Students are asked to give overall ratings of both their instructor and their course. In addition, they are asked to rate specific characteristics of the instructor (e.g., knowledge, fairness, helpfulness) and of the course (e.g., organization, difficulty, informative). Mean ratings are then computed across all students and for each rated item, as well as across all rated items. These mean ratings are often used to evaluate a professor's teaching effectiveness by comparing them with ratings received by other professors in the department or in the faculty (Uttl et al., 2017). Whereas in 1973 only 29% of colleges collected SETs, this practice increased to 68% in 1983 and to 86% in 1993 (Seldin, 1998). A survey conducted in 2010 indicated that SETs were collected in 94% of colleges, that nearly all deans declared that classroom teaching was a major part of the performance evaluation of their faculty, and that SETs were usually their main source of information about the quality of classroom teaching (Miller & Seldin, 2014).

The use of SETs as a basis for decisions on promotion and tenure is justified only if SETs are a valid measure of teaching effectiveness and student learning. But they are not (e.g., Boring et al., 2016; Uttl et al., 2017). They are most likely a reflection of students' satisfaction with a course, which can be influenced by many factors that are unrelated to teaching effectiveness (Freishtat, 2016). In the first section of this article, evidence is presented that SETs are not a valid measure of student learning. In the second section a process model is proposed, showing how SETs encourage poor teaching and cause grade inflation. The third section reviews evidence for the processes assumed by the process model. The fourth section illustrates the extent of grade inflation and discusses the dark side of this inflationary process. Finally,

**CONTACT** Wolfgang Stroebe wolfgang.stroebe@gmail.com Department of Social and Organizational Psychology, University of Groningen, Grote, Kruisstraat 2/1, Groningen, 9712TS, Netherlands.

conclusions about a responsible use of SETs are discussed.

## On the validity of student evaluations of teaching

SETs have a great deal of face validity. After all, students take many courses and should therefore be able to judge whether a particular instructor is an effective teacher. Supporting this assumption is the evidence that students' course evaluations are positively correlated with the grades they received in those courses. This can be interpreted as supportive if we make the following assumptions: (a) Students learn more from good teachers, (b) course grades are a good measure of learning, and (c) students are able to evaluate the quality of the teaching of their instructor. However, because all students in a class are exposed to the same instructor, one could wonder if students who learn more also perceive him or her as a better teacher. Another more fundamental critique is that the correlation between students' course grades and their evaluation of a teacher might merely reflect bias. Although students receive their course grades at the end of a course, they typically already have a good idea what to expect. A bias interpretation would assume that students who expect to receive a good grade evaluate a teacher more positively than students who expect to receive a poor grade.

## Multisection studies of student evaluations of teaching

One way to distinguish between these two interpretations is to test whether the relationship between course grades and SETs would be maintained if the correlation of the average SET scores of a set of classes were correlated with the grade point average (GPA) of those classes. If SETs reflect teaching effectiveness rather than bias, and if course grades reflect learning, then the average SET of a set of classes should be positively correlated with the average GPA of these classes. This prediction has been tested with *multisection courses*. These are courses that are split into a number of parallel sections, each taught by a different instructor. An optimal multisection course should meet the following requirements: (a) It should have many sections in which the same material is being taught, (b) each section should be taught by a different instructor, (c) students should be randomly assigned to these sections to avoid self-selection, (d) all sections should be assessed with the same centrally administered exam, and (e) SETs should be administered either just before or with the exam.

In a study by Boring et al. (2016) to be described later, all 1st-year students took the same mandatory courses. In each course, main lectures were given by a professor to approximately 900 students. Courses were divided into sections of 10–24 students taught by instructors. The final exam, taken by all students, was written by the course professor. If students rated sections in which they learned a great deal more positively than sections in which they learned little (and if the exam was a valid measure of student learning), average-section SETs should correlate positively with average-section grades (GPA). As it is often impossible to assign students randomly to sections, some multisection studies correct final grades by indicators of prior learning or ability (e.g., overall GPA, SAT scores). The multisection design is considered the gold standard in research on the validity of SETs (e.g., Abrami et al., 1990; Cohen, 1981, 1983; Feldman, 1989).

Early meta-analyses of such multisection studies (e.g., Cohen, 1981; Feldman, 1989) concluded that there was a moderately positive correlation between SET averages and GPAs. Most influential was the meta-analysis of Cohen, which was based on 68 multisection studies of which 67 provided useful data. The average correlation between overall instructor ratings and GPA was $r = .43$, a moderately large effect. In a critique of Cohen's (1981) meta-analysis, as well as the later meta-analysis by Feldman (1989), Uttl et al. (2017) pointed out that the number of sections included in most of their multisection studies was rather small and that these small studies often had extremely high correlations. For example, more than one third of Cohen's multisection studies had 10 or fewer sections (Uttl et al., 2017). This would have been less of a problem if these authors had corrected for sample size. However, as Uttl et al. (2017) criticized, these meta-analyses gave the same weight to all studies, independent of their sample size. Cohen (1981) denied that this could be a problem, even though he reported that an analysis of studies that used at least 20 or more sections resulted in a lower average correlation of $r = .37$. As Uttl et al. (2017) added, the correlation is reduced to $r = .27$ if one bases one's analysis only on multisection studies with 30 or more sections.

A more recent meta-analysis by Clayson (2009) reported that the unweighted average correlation between SETs and learning was $r = .33$, whereas the correlation weighted by sample size was only $r = .13$.

Of interest, Clayson also reported a correlation of $r = .48$ between the SET–learning correlation and the year of publication, with effects being highest in early studies. He also found that the correlation between number of sections and the SET–learning correlation was $r = -.37$, with studies having few sections achieving higher correlations. However, Clayson's meta-analysis is problematic, because he included Cohen's meta-analysis as one of his multisection studies with a SET–learning correlation of $r = .41$ and 35 sections. As Uttl et al. (2017) remarked, "We cannot think of any reason mixing the meta-analysis estimated $r$ with multisection studies' $r$ to conduct another meta-analysis of multi-section $r$s" (p. 31).

Uttl et al. (2017) conducted a larger meta-analysis based on 97 multisection studies reported in 51 articles, which they identified after a systematic literature search. Again, studies with few sections often reported extremely high correlations. In fact, the correlation between number of sections and SET–learning correlation was $r = -.35$. When they controlled for small study effects, the estimated SET–learning correlation was $r = .12$. Including only studies that controlled for prior knowledge/ability reduced the SET–learning correlations to $r = -.06$. Uttl et al. (2017) concluded that "multisection studies do not support the claims that students learn more from highly rated professors" (p. 35).

A literature search on Web of Science and Scopus (search term: "student evaluation of teaching") revealed only one recent multisection study published after Uttl et al., (2017) stopped their literature search (Boring et al., 2016; also discussed in Boring, 2017). This study uses data from European (i.e., French) students and is based on 23,000 SETs from 4,423 first-year students in 1,177 sections taught by 379 instructors. The data are particularly interesting because there was no self-selection into sections and responses to SETs were compulsory so that there was a near 100% response rate. The average correlation between SETs and final exam scores was $r = .04$.

How is it possible that it took 40 years to discover that SETs are unrelated to student learning? One reason could be that the majority of studies had been conducted before 1981. As Clayson (2009) reported, the size of SET–learning correlations were strongly correlated with year of publication. When Uttl et al. (2019) looked in more detail at this association, they found that the 69 studies published prior to 1981 yielded an average correlation of $r = .31$, whereas the 28 studies published in or after 1981 had an average correlation of $r = .06$.

This is a substantial difference, and one wonders how such a difference can be explained. Uttl et al. (2019) offered conflict of interest (COI) as an explanation: Many of the early studies were published by researchers who either worked for or owned corporations that sell SET systems and thus had a (financial) interest in finding substantial SET–learning correlations. There is ample evidence from pharmaceutical research that funding sources influence research outcomes (e.g., Bekelman et al., 2003; Lundh et al., 2018). Vartanian et al. (2007) even reported that findings of the effects of sugar-containing soft drinks on body weight were influenced by the funding source: Sugar-containing soft drinks had less impact on weight when the study was paid for by the producer rather than a neutral source.

Although these effects of COI are well known, the finding that COI also influenced research on the effectiveness of SETs as measure of teaching quality is new. However, it is not surprising. After all, it can be assumed that a researcher who either works for or owns a firm that sells SET systems would prefer to find a high rather than a low correlation between SET and teaching effectiveness. As Upton Sinclair (1934/1940) once wrote, "It is difficult to get a man to understand something, when his salary depends on his not understanding it" (p. 109). The evidence presented by Uttl et al. (2019) is persuasive. They found that SET–learning correlations were much larger when at least one author of a study was associated with an SET corporation. Whereas the correlation was $r = .58$ for the 15 studies with a corporate COI, the correlation was $r = .18$ for 82 studies of authors without corporate interests (Figure 1).

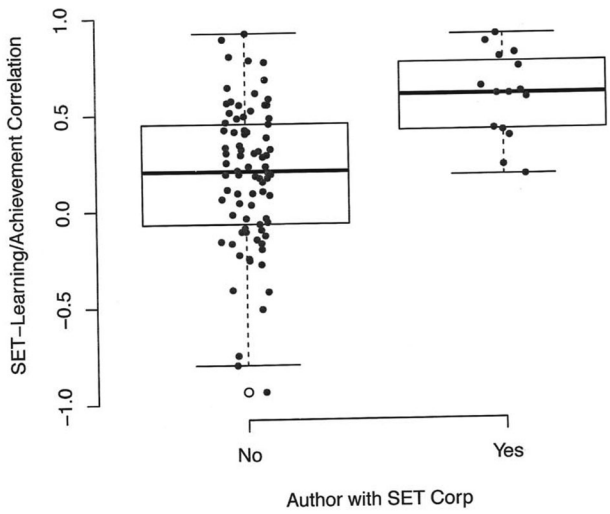


**Figure 1.** SET–learning correlations as a function of conflict of interest (from Figure 2B of Uttl, Cnudde & White, with permission of the authors).

In an analysis of the effect of COI on outcome of pharmaceutical research, Sismondo (2008) suggested that such conflicts might not operate on a conscious level but that accepting funds from industry creates a gift relationship between the investigator and the sponsor, in which the sponsor might feel a need to reciprocate. In psychological research, there are subtle ways in which a researcher's motivation due to COI can unknowingly influence study findings. Researchers may be less motivated—and therefore less likely—to scrutinize results that support their hypotheses than findings that are inconsistent. An example that suggests lack of motivation to scrutinize inconsistencies—mentioned by Uttl et al. (2017)—is in an article by Abrami et al. (1988). In comparing early meta-analyses, these authors observed disagreements between data extracted by Cohen (1981) and McCallum (1984) from the same studies but did not follow this up. When Uttl et al. (2017) checked the extracted data for accuracy in the original studies, they discovered that a large proportion of the McCallum data was simply incorrect.

## The influence of teaching-irrelevant factors on student evaluations of teaching

SETs are influenced by numerous variables that are unrelated to teaching effectiveness, such as gender and race. Other variables, such as likeability of an instructor—which seem to be irrelevant—could be related to effectiveness. For example, the instructor may be liked because he or she is accessible and helpful to students. Unfortunately, research on these characteristics has exclusively focused on demonstrating the (direct) relationship with course ratings or instructor ratings. Rarely have researchers tried to assess theoretically plausible variables that might mediate the association between such factors and overall ratings of instructors or courses. Because course teachers are communicators, variables known to increase the effectiveness of a communicator (e.g., perceived expertise, likeability, power) are likely to increase an instructor's teaching effectiveness. Thus, characteristics that seem to have no direct relationship with teaching effectiveness might have indirect effects through related processes that are associated with teaching effectiveness.

One of the most blatantly irrelevant characteristics that has been repeatedly shown to be strongly associated with SETs is the physical attractiveness of instructors. Because institutional SETs do not assess physical attractiveness, most of the evidence that attractiveness influences students' evaluation of teachers has been provided by studies using the RateMyProfessors.com (RMP) website. An added advantage of RMP is that, unlike institutional SETs, this information is publicly available. Because the correspondence of RMP ratings to institutional SETs has often been questioned (e.g., Legg & Wilson, 2012; Murray & Zdravkovic, 2016), I discuss this issue before reviewing some RMP findings.

### The correspondence of RMP ratings to institutional SETs

Established in 1999, RMP is a popular website where students can evaluate their professors on four dimensions: helpfulness, clarity, easiness, and "hotness." Helpfulness and clarity are combined into an indicator of quality of teaching.[1] Most studies comparing evaluations of instructors on RMP to institutional SETs have found substantial correlations (e.g., Brown et al., 2009; Colardarci & Kornfield, 2007; Sonntag et al., 2009; Timmerman, 2008) suggesting a fair degree of equivalence. For example, Timmerman (2008), who identified 1,002 professors at the University of California, San Diego, with both RMP and SET ratings, found a correlation of $r = .66$ between the percentage of students who would recommend an instructor and RMP ratings of overall quality. The correlation between RMP overall quality and the percentage of students' recommending a class was $r = .51$ and between RMP and self-reported learning was $r = .57$. Similar findings were reported by Sonntag et al. (2009) based on RMP ratings of 104 Lander University professors. The SET ratings used were ratings of professors and of classes as excellent on 5-point scales. The correlation of RMP overall quality ratings with SET ratings of instructor excellence was $r = .69$ and with class excellence was $r = .60$. Similar correlations were reported by Brown et al. (2009) in a study based on 312 Brooklyn College instructors and by Colardarci and Kornfield (2007) in a study based on 283 instructors at the University of Maine. The size of these correlations is particularly surprising, if one considers that RMP ratings are typically based on much smaller numbers of students than those who participated in the SET evaluations.

Since Legg and Wilson (2012) published a paper with the suggestive title "RateMyProfessors.com Offers Biased Evaluations," a brief review of their study is warranted. They collected three sets of RMP ratings of 25 professors willing to participate in the study. The first set comprised ratings of previous classes given by

those professors. The second comprised ratings of RMP items embedded in an SET administered during classes. These students were later (probably at the end of the semester) asked to rate the class on RMP. The main finding was that in-class and end-of-semester evaluations of clarity were slightly higher (by approximately 0.5 points) than clarity ratings collected on previous classes. Professors were also rated less helpful and easier in classes given before the study. One explanation for these differences could be that these instructors were on their best behavior during the study.

A critical study by Murray and Zdravkovic (2016) is similarly unconvincing. These authors compared RMP ratings of instructors with a 12-item scale that

> addressed instructor and course aspects related to whether the professor (a) enjoys teaching, (b) is well organized, (c) is friendly and considerate of students, (d) makes challenging assignments, (e) is available to provide extra help and (f) is enthusiastic about the course material. (p. 141)

These items were individually weighed by each study participant according to perceived importance. The authors reported that the evaluation of teaching was higher ($M = 3.80$) when measured with the six-item scale rather than the RMP scale ($M = 3.56$). The authors present no evidence that their complex index constituted a more valid measure of teaching effectiveness than the RMP. Furthermore, the difference between their rather complex measure and the simple RMP ratings is minor.

### Physical attractiveness

Physically attractive (i.e., "hot") instructors receive more positive teaching ratings than their less attractive colleagues (e.g., Boehmer & Wood, 2017; Felton et al., 2008; Fisher et al., 2019; Freng & Webber, 2009; Hamermesh & Parker, 2005); Johnson & Crews, 2013; Riniolo et al., 2006; Rosen, 2018; Wolbring & Riordan, 2016). In an early RMP study that used data for all the professors in the United States and Canada who had at least 20 student ratings (6,851 professors from 369 institutions), Felton et al. (2008) reported a correlation of $r = .64$ between "hotness" and "quality."

A more recent RMP study by Rosen (2018), which was based on 7,882,980 ratings of 190,006 professors from 4522 U.S. colleges and universities, who had a minimum of 20 ratings, replicated the strong association between hotness and quality. Hardly any professor with a quality rating below 2.5 was considered hot, compared to 70% of professors with a perfect quality rating. Similarly strong effects were reported by Riniolo et al. (2006). Other RMP studies found weaker associations. For example, Freng and Webber (2009) reported a correlation of $r = .37$ and Johnson and Crews (2013) of $r = .16$.

The fact that these effects of physical attractiveness are not limited to RMP ratings has been demonstrated in two studies that related perceived attractiveness of instructors (based on ratings of photographs from websites) to their end-of-semester SETs. Based on a study of 94 instructors, Hamermesh and Parker (2005) reported that "the instructional rating varies by nearly two standard deviations between the worst- and best-looking instructors in the sample" (p. 372). In a study conducted in Germany with 125 instructors, Wolbring and Riordan (2016) found a somewhat weaker association of physical attractiveness with SETs. However, SET ratings also predicted absenteeism from class: Students missed slightly fewer classes of physically attractive instructors. It made no difference whether students and instructors were of the same or opposite sex.

Wolbring and Riordan (2016) replicated their association of physical attractiveness and teaching evaluation in an experimental study in which students were exposed to pictures (attached to a CV) of either an attractive or a somewhat less attractive man or woman. Students then had to listen to—and evaluate—a lecture supposedly given by this stimulus person. Again, physical attractiveness influenced ratings of the quality of that lecture. Finally, Ambady and Rosenthal (1993) reported that ratings (based on a 30-s silent film clip) of the physical attractiveness of 13 college teachers by two female undergraduates correlated $r = .32$ with students' end-of-term ratings of the quality of courses taught by these instructors.

### Likeability of instructor

One way that physical attractiveness might increase the perceived effectiveness of an instructor is through its association with likeability. Although likeability is a broader concept, physical attractiveness is a major determinant of likeability. A study that asked 861 students from a U.S. university to rate the physical attractiveness of two instructors of classes they were currently taking reported a correlation of $r = .47$ with liking (Gurung & Vespia, 2007). The correlation between liking their instructor and enjoyment of the class was $r = .80$. The correlation with attendance was also positive ($r = .30$), suggesting that students will attend a class more regularly if they like an instructor.

Similar findings were reported by Feistauer and Richter (2018) based on a sample of 260 students rating 26 instructors who taught psychology lectures or seminars at a German university. Feistauer and Richter studied the association of likeability with the evaluation of instructors, whose seminars or lectures on psychology the students had attended. Likeability was rated twice with a single-item measure administered during the first 10 min of the first session and before the end of the semester. The overall performance of instructors was assessed with a single-item scale toward the end of the semester. The two liking ratings were moderately correlated (.55–.58). Liking measured at the beginning of the semester accounted for 9.4% of the total variance in overall ratings of the instructor for lectures and 20.5% of the variance for seminars. Liking at the end of the semester (i.e., at the time of the evaluation) accounted for 36.5% of the variance for lectures and 54.7% for seminars.

A study by Delucchi (2000) provides some indications about the characteristics of instructors that influence likability. Based on a factor analysis of SET ratings, Delucchi identified a likeability index that consisted of some of the following evaluations. A likeable instructor had good rapport with students, was easy to talk to, seemed enthusiastic about the subject matter, and created a feeling of community among students. These findings are consistent with findings reported by Feistauer and Richter (2018), who had also administered the Reysen (2005) Likability Scale at Time 1 to establish the validity of their single-item measure. The Reysen scale assessed whether the instructor is perceived by students as friendly, likeable, warmhearted, approachable, knowledgeable, physically attractive, and similar to them. The Reysen scale correlated $r = .98$ in lectures and $r = .89$ in seminars with the single-item measure given at the same time and $r = .46$ and $r = .45$ with the single-item measure given at the end of the semester. Because likability assessed during the first 10 min of the first class meeting is likely to be mainly based on physical appearance, these (initially) high correlations suggest that the Reysen scale perfectly captures the stereotype people hold of physically attractive others. The finding that initial ratings of liking were only moderately correlated with liking at the end of the semester and that initial liking accounted for much less variance than later liking in the end-of-semester overall rating of instructors is consistent with the conclusions of the classic meta-analysis of the attractiveness stereotype that the effect is weakened when individuals receive individuating information (Eagly et al., 1991). In the course of the semesters, students might have discovered that their instructors were not quite as warmhearted, approachable, and knowledgeable as they had expected them to be based on their first impression.

Likeability appears to be the result of characteristics that are certainly part of good teaching, such as approachability, friendliness, and knowledgeableness. The finding that physical attractiveness is correlated with these characteristics suggests the possibility of a halo effect: Students assume that a physically attractive instructor also possesses these traits. This interpretation would explain why physical attractiveness is related to class attendance, or why "hotness" in RMP studies is always highly correlated with perceived helpfulness. It would finally explain why, in the study of Feistauer and Richter (2018), students decided after being acquainted for 10 min that their instructor was warmhearted, approachable, knowledgeable, and similar to them.

### *Prior subject interest*

One would expect that students rate courses on topics in which they are interested more positively than courses of little interest. Surprisingly the evidence is mixed. Olivares (2001) found prior interest—measured at the beginning of the first class meeting—unrelated to global rating the course. In contrast, Griffin (2004) reported a moderate correlation between initial interest and teacher ratings ($r = .37$) and an even higher correlation with ratings of a course ($r = .50$). Feistauer and Richter (2018) also reported a weak but statistically reliable association.

Students will often have false expectations about the content of a course and how interesting the material will be to them. This is particularly likely in psychology courses, in which beginning students often have unrealistic expectations of what they will learn. Although many expect that they will learn how to know people and how to help them, they find themselves exposed to learning theories and statistics. It is therefore interesting that Olivares (2001), who failed to find initial interest related to instructor ratings, found that interest change assessed at the end of the semester moderately correlated with student ratings of their instructors ($r = .42$). There must have been quite a bit of change happening in these classes, as the correlation between initial interest and interest change was only $r = .21$. It seems plausible that interest in the content of a course influences enjoyment of the course and therefore also ratings of the instructor. Students who are totally uninterested in a course are unlikely

to give an instructor very positive ratings. Arousing students' interest in the subject that one is teaching is certainly a characteristic of teaching ability. Unfortunately, it is easier if one teaches social psychology rather than statistics.

### Minority status

Because it is implausible that teachers belonging to a minority are generally less able instructors than majority teachers, minority status should have no impact on SETs. And yet, most studies show that non-White instructors receive SET ratings that are lower than that of their White colleagues (e.g., McPherson & Jewell, 2007; Reid, 2010; Smith, 2007). In a study based on data from 24 consecutive semesters, comprising 280 graduate classes taught by 22 instructors, White instructors received higher SET scores than their non-White colleagues (McPherson & Jewell, 2007). Similar results were reported by Smith (2007) based on a much larger sample of minority faculty members from a college of education situated in the southern United States. Again, Black faculty members received lower evaluations of teaching effectiveness than their White colleagues, with ratings of other minority faculty (Latinos, Asians, Native Americans) falling in between. Hamermesh and Parker (2005) also reported that minority faculty members and non-native English speakers received substantially lower teaching ratings than majority faculty members and native English speakers.

A similar pattern emerged from an RMP study based on ratings of 5,630 faculties of liberal arts colleges (Reid, 2010). Because RMP does not list race of instructors, Reid had a multiracial group of students decide on race based on photographs of these faculty members. Racial minority faculty members were rated less favorably than White faculty members on quality, helpfulness, and clarity. However, they received more positive ratings on easiness. These differences were mainly due to lower quality ratings received by Black faculty members. The most likely determinant of the lower ratings received by minority status teachers is prejudice. Prejudiced individuals would tend to perceive minority faculty members as less intelligent and possessing less expertise.

### Gender

Evidence for gender differences is less consistent. However, if gender differences are found, it is mostly female instructors, who receive lower ratings on teaching quality. The RMP study of Rosen (2018) found that women received slightly lower scores on quality. Such gender differences were also reported in an RMP study of Boehmer and Wood (2017) and by Arceo-Gomez and Campos-Vazquez (2019) in a study based on a Mexican internet site (MisProfesores.com) but not in RMP studies of Reid (2010) and Stuber et al. (2009).

Three large multisection studies based on institutional SETs also reported gender differences favoring men (Boring, 2017; Hamermesh & Parker, 2005; Mengel et al., 2019). Using the dataset of Boring et al. (2016) described earlier, Boring (2017) found that both male and female students gave male professors slightly higher ratings on overall satisfaction, with the difference somewhat less marked for female students. However, these differences were minimal. In contrast, Hamermesh and Parker (2005) reported that female instructors received ratings that were nearly half a standard deviation lower than those of their male colleagues. Finally, in a study based on nearly 20,000 student evaluations conducted at the School of Business and Economics of Maastricht University (Netherlands), Mengel et al. (2019) found that women received systematically lower ratings than their male colleagues, even though neither grades nor students' study hours were affected by the gender of their instructor. These gender differences were particularly marked for courses with mathematical content.

This last finding raises the possibility that gender differences could be moderated by discipline (e.g., science vs. humanities). However, the evidence is inconsistent. An RMP study by Fisher et al. (2019) found more gender discrimination in fields such as engineering, business/economics, and computer science than in English, history, and philosophy—even when they controlled for percentage of female staff in these departments. However, Stuber et al. (2009) did not find such differences. These authors defined the hard sciences, engineering, mathematics, and business as traditionally male areas and arts and humanities as traditionally female. Although they replicated the typical finding that instructors teaching in those traditionally male fields were rated more negatively than those teaching humanities or arts, there was no interaction with gender: Women were not more penalized than men.

Could the lower ratings typically received by female instructors reflect lower teaching ability? A highly-cited study by MacNell et al. (2015) appeared to rule out this interpretation. These authors conducted an experiment in which they manipulated the *perceived*

gender of instructors. Two assistant instructors (one male and one female) in an online class each operated under two different gender identities. Regardless of actual gender, male-identity teachers received higher evaluation on professionalism, promptness, fairness, respectfulness, giving praise, and enthusiasm. However, Uttl and Violo (2020) questioned these findings on several accounts: They argued that one could hardly generalize to all male or female instructors based on findings with only two individuals. Furthermore, the sample of students in each condition was rather small, ranging from eight to 12 individuals. But most critically, there were three outliers; they gave the lowest ratings on all SET items in the two female conditions. If one removed these outliers, the gender difference disappeared. Instead, students rated the actual female instructor higher than the male instructor, regardless of perceived gender.

Another study that varied gender was conducted by Mitchell and Martin (2018). Dr. Kristina Mitchell or Dr. Jonathan Martin, who are faculty members at different colleges, gave an online course of identical content at their respective colleges.[2] The course was rated more positively when attributed to a male instructor rather than a female instructor. Because the courses were given to different student populations, gender of instructor was confounded with the student population rating the courses. Furthermore, because Drs. Mitchell and Martin were likely to be known to their students, it seems possible that preexisting attitudes toward these instructors rather than mere gender information could have influenced these ratings. I hasten to add that the fact that we cannot rule out this hypothesis does not mean that it is plausible.

### *Discipline*

Whereas people have no choice regarding their physical attractiveness, race, or gender, they typically can choose their discipline. There is ample evidence that if they want to receive positive ratings for their teaching, they should choose humanities or languages rather than mathematics, engineering, or computer science. Using Educational Testing Service data from 238,471 classes, Centra (2009) found that compared with classes in humanities (English, history, languages), natural science classes (mathematics, engineering, and computer science) were rated 0.30 standard deviations lower. Consistent with this, Felton et al. (2008) reported from their RMP study that the departments with the highest quality ratings were languages, sociology, and political science, with engineering, computer science, and chemistry as the lowest ranking departments. A similar difference was found by Uttl and Smibert (2017) in a comparison of SETs of English and Math classes at New York University. The average SET of English classes was 4.29; the math average was 3.68. Consistent with the fact that SETs are used for personnel decisions, Uttl and Smibert (2017) found that professors teaching quantitative courses were also less likely than their colleagues in English departments to be tenured, promoted, or given merit pay.

## Conclusion

Student evaluations of teaching do not measure teaching effectiveness. Furthermore, they are influenced by several factors unrelated to teaching quality such as minority status, foreign accent and gender of instructors, and the discipline they are teaching. This makes their use by university administrators in decisions about hiring, salary increases, and promotions unfair and potentially illegal. In fact, in a recent arbitration decision regarding a conflict between the Canadian Ryerson University and its faculty association (Ryerson University v Ryerson Faculty Association, 2018, CanLII 58446), an arbitrator argued that "insofar as assessing teaching effectiveness is concerned – especially in the context of tenure and promotion—SETs are imperfect at best and downright biased and unreliable at worst." He decreed that the university should no longer use SETs as a measure of teaching effectiveness in promotion and tenure decisions.

## The biasing effects of student evaluations of teaching: a process model

In this section, I argue that this misuse of SETs in decision about hiring, salary increases, and promotions could be the cause of bad teaching and grade inflation. University administrators place great emphasis on good SETs, which makes getting good SETs highly important for instructors. This provides students with an effective tool to influence the type of teaching they receive. This would be no problem if SETs were a valid indicator of teaching effectiveness and if students were motivated only by a need to learn. However, because SETs are unrelated to teaching effectiveness and because—as I show—a sizeable proportion of students does not seem motivated to work hard and to learn a great deal (e.g., Chiu et al., 2019; Felton et al., 2008; Rosen, 2018) and prefers

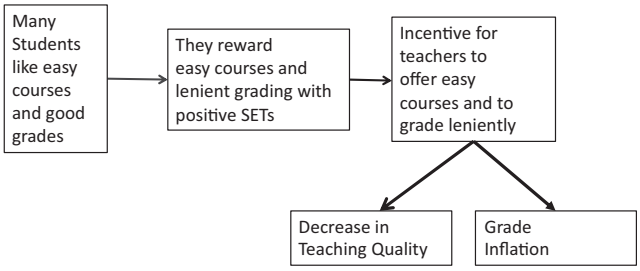


**Figure 2.** The biasing effect of SETs on teaching quality and grades: A process model.

easy courses and lenient grading, their feedback—if accepted—is likely to reduce the quality of teaching. To describe this process, a model is presented that is based on empirically supported assumptions about the motivation of a majority of students. After presenting the model, I review evidence that supports each of the processes assumed by the model.

This analysis starts with the assumption that all students like getting good course grades. Many of these students would also prefer not to spend too much time on reading, writing, and other class preparation (Figure 2; Chiu et al., 2019; Felton et al., 2008; Rosen, 2018). One way to achieve these normally conflicting goals is to take classes with teachers who are known to grade leniently and not require too much coursework (Johnson, 2003; Sabot & Wakeman-Linn, 1991). As a result, many students prefer easy classes that promise good grades to challenging courses, where hard work is required and good grades are not a certainty (e.g., Bar et al., 2009; Johnson, 2003; Sabot & Wakeman-Linn, 1991). The students therefore give the reward of good SETs to teachers who grade leniently (e.g., Anderson et al., 1975; Greenwald & Gillmore, 1997a; Griffin, 2004; Olivares, 2001) and do not require too much work (e.g., Felton et al., 2008; Rosen, 2018).

Teachers not only like their courses to receive good SETs but also know that good SETs are important for promotion and merit increases, which creates an incentive to offer easy courses and to grade leniently (Birnbaum, 2000; Keng, 2018; Moore & Trahan, 1998; Ryan et al., 1980; Simpson & Siguaw, 2000). The stronger the incentive, the stricter the comparison levels used in a department (e.g., Are instructors required to score above the mean of the department, or even above the 70th percentile? Is the comparison group the department, the faculty, or the whole university?). The most proximal result of this process is that SETs often reward poor teaching and lenient grading. The more distal effect of this race to the bottom is grade inflation.

This model offers an explanation for the great paradox of American university education—namely, that GPAs have increased for decades (Rojstaczer, 2015), yet university students have not become more hardworking or better qualified for college.[3] On the contrary, SAT scores show a downward trend (e.g., Adams, 2015; Washington Post, 2015), and students spend less time on academic pursuits today than they did a few decades ago (e.g., Arum & Roksa, 2011; Babcock & Marks, 2011). There is even evidence to suggest that a college education results in a lower gain in critical thinking, complex reasoning, and writing skills today than it did several decades ago (Arum & Roksa, 2011; Pascarella et al., 2011). According to the simple process model suggested here, the widespread use of SETs is a major cause of these effects.

## Causes of bias: the evidence

This section discusses how students' preference for good grades and easy courses biases SETs and influences course choices. Building on Stroebe (2016), evidence for four propositions is presented: (a) Students reward good grades with positive SETs, (b) students reward easy courses with positive SETs, (c) students choose courses that promise good grades, and (d) instructors want (need) good SETs.

### *Students reward good grades with positive SETs*

Anderson et al. (1975) demonstrated this bias with a simple study. They assessed both grade expectations and SET twice in a class, once at the end of the first session of a class and again in the last session, but before the final exam. They then divided their participants into those whose grade expectations became worse, improved, or remained the same. In support of a bias interpretation, the overall ratings of instructor and course improved with improving grade expectations and decreased substantially with decreasing grade expectations. The same pattern was reported by Clayson et al. (2006) in a study of 499 students of undergraduate business courses. At Week 10 and again at Week 16, students were asked to evaluate their instructors and their expected grades. In line with the findings of Anderson et al., Clayson et al. found changes in expected grades to be associated with corresponding changes in students' evaluation of their instructor.

A similar pattern was observed when Wellesley College introduced a grade ceiling for lenient-grading departments (Butcher et al., 2014). In the early 2000s,

the faculty and administration at Wellesley decided that the credibility of the institution was threatened by grade inflation. As is generally the case, grade inflation was mainly a problem in the humanities and social sciences but did not affect science departments. The college instituted the rule that the average grade must not exceed a B+ (3.33) in introductory and intermediate courses. Although the grade ceiling was effective in lowering average grades, it had the unexpected side effect that it also lowered faculty ratings. The percentage of students strongly recommending their professor decreased by 5%, and there was an increase in *neutral* and *do not recommend* categories. As Butcher et al. (2014) concluded, "The results strongly indicate that students were less pleased with their instructors, when the grading policy lowered average grades" (p. 200). In 2019, the college rescinded this grading policy (https://www.wellesley.edu/registrar/grading/grading_policy). Similar findings were reported from a study conducted at a large state university where grade ceilings were introduced in required business school classes (Gorry, 2017). The grade ceiling was set at 2.8 in introductory courses and at 3.2 in intermediate courses. Both ceilings were effective in lowering average grades. But while the 2.8 ceiling significantly lowered course evaluations and also increased the number of withdrawals, the effect of the higher ceiling on course evaluation was much smaller.

Further support for the hypothesis that SETs are biased by grade expectations comes from a study by Greenwald and Gillmore (1997a). These researchers added three items to an SET that were fairly unrelated to teaching quality. Students had to rate the legibility of the instructor's handwriting, the audibility of his or her voice, and the quality of classroom facilities. In line with a bias interpretation, they found a positive correlation between expected grades and the positiveness of ratings on these items. Because all students heard the same voice, read the same writing, and worked in the same classroom, the finding of this within-class correlation suggests that grade expectations biased these evaluations. Furthermore, this correlation could not be observed for between-class analyses, which further supports the assumption that these qualities were unrelated to teaching effectiveness.

Even more direct evidence for a bias explanation has been provided by Olivares (2001), who related perceived grading leniency to SETs. In a study based on 149 students and seven sections of two undergraduate courses taught by the same instructor, perceived grading leniency was measured directly with the following question: "Compared to all other college instructors you have had, how would you rate this instructor's grading?" The response scale ranged from *much harder/strict grader* to *much easier/lenient grader*. In measures taken at the end of the semester, grading leniency correlated with both the global rating of the instructor ($r = .45$) and the multiple-items SET scale ($r = .45$). A somewhat lower correlation was reported by Griffin (2004) in a study based on 754 undergraduate students enrolled in 39 education courses. Grading leniency was assessed with the statement, "This instructor is a lenient/easy grader (*strongly agree* to *strongly disagree*)." The correlation with the average of all SET ratings was $r = .23$ and thus lower than in the study of Olivares (2001). However, in both studies, perceived grading leniency was positively correlated with instructor evaluation, suggesting that instructors should be able to improve their teaching ratings by grading leniently.

If instructors do not want to grade leniently or lower their course requirements, they can use other strategies to improve their SETs. In a study by Youmans and Jee (2007), half of a set of classes were treated with chocolate bars on the day they had to respond to the SET. The person giving the chocolate bars was independent of the class instructor and (allegedly) had these bars over from another function. The average SET score was higher for the classes receiving chocolate (4.07 vs. 3.85). A similar difference in ratings was observed in a study where half of the classes were given cookies during a first session and then had to evaluate the teaching quality of that session (Hessler et al., 2018).

### Students reward easy courses with good SETs

Support for this proposition comes mainly from studies of the RMP website described earlier. On that website, easiness is clearly defined in terms of lenient grading. Students are instructed to ask themselves, "How easy are the classes this professor teaches? Is it possible to get an A without much work"? Easiness is consistently found to be a strong predictor of the evaluation of teaching effectiveness in RMP studies (Boehmer & Wood, 2017; Felton et al., 2008; Johnson & Crews, 2013; Rosen, 2018). In the study by Felton et al. (2008), described earlier, quality and easiness correlated at $r = .62$. Students rated courses that enabled them to get excellent grades without doing much work more positively than courses that required a great deal of time and effort. Although one could

argue that easiness might be a consequence of quality, with brilliant teachers making even the most difficult material easy to understand, this interpretation would be inconsistent with the way easiness is defined. Furthermore, RMP allows students to write comments about a professor to justify their ratings and "professors with high 'Easiness' scores usually received comments regarding a low workload and high grades" (Felton et al., 2008, p. 40). These findings were replicated by Rosen (2018) with a correlation between overall quality and easiness of $r = .61$ that is practically identical to the correlation reported by Felton et al. RMP studies by Boehmer and Wood (2017) and Johnson and Crews (2013) reported similar correlations (.66 and .62, respectively). A somewhat lower correlation of $r = .35$ between easiness ratings and the percentage of students recommending an instructor was reported by Timmerman (2008) in the RMP study described earlier.

The association between easiness and quality was found to be somewhat weaker at top universities in another RMP study using data of 85,306 professors from 3,799 colleges and universities (Chiu et al., 2019). These researchers used the 2016 Forbes classification of top 200 U.S. colleges and universities (http://www.forbes.com/top-colleges) to classify the colleges and universities of their sample into top or nontop institutions. Although they replicated the strong main effect that easy courses were rated as having a higher quality, they also found easiness to interact with their college classification: Students at top colleges evaluated easy courses slightly less positively on quality than students from nontop colleges (4.0 vs. 4.13, respectively) but rated difficult courses slightly more positively (3.37 vs. 3.27, respectively). The most plausible interpretation of these (minor) effects is that top colleges attract students who are more willing to work hard, which enables them to cope better with more difficult courses.

Because some researchers doubt that RMP ratings are equivalent to ratings on university SETs (e.g., Legg & Wilson, 2012; Murray & Zdravkovic, 2016), it is important to note that this equivalence is not essential for the argument presented here. The fact that "easiness" has been found to be highly correlated with evaluations of the "quality" of a course in students' ratings is sufficient evidence that the two dimensions are closely and positively associated in the minds of these students.

## Students choose courses that promise good grades

According to the revealed preference theory pioneered by the economist Samuelson (1948), the preferences of consumers can be revealed by their purchasing behavior. Because this theory should also apply to course choices, the preference of students for courses of teachers who grade leniently should be revealed by their course choice. In an early study, Sabot and Wakeman-Linn (1991) assessed the likelihood that students took a second course in a department as a function of the grades they received in their first course. Of students who did not intend to major in economics but had taken a course in that department, the probability of taking a second course was 18% lower if they received a B rather than an A and 28% lower if they received a C in an introductory course in economics.

Similarly, in a longitudinal study that provided information about the extent to which students informed themselves about the average grade of courses taught in the previous semesters, Johnson (2003) found that this information influenced students' future choices. If a student had a choice between courses taught by two instructors, one course having a GPA of A– and the other having a GPA of B, the odds that a student would choose the first course over the second were 2 to 1. Finally, in a study of persistence toward graduating in the physical and life sciences, Ost (2010) found that students who received higher grades in their nonscience courses than their science courses were more likely to transfer out of the sciences than were students who received higher grades in their chosen science field.

Further evidence for students' "revealed preference" for leniently graded courses comes from a study at Cornell University (Bar et al., 2009). This university decided as of 1998 to publish median grades for all courses on a website and to mention them in students' transcripts. The university hoped that a more "accurate recognition of performance may encourage students to take courses in which the median grade is relatively low" (p. 94). This hope was not supported. In fact, the proportion of courses with an A median increased by 16% after the introduction of this website. But even more striking, the proportion of students who enrolled in such courses increased by 42%. However, this increase was mainly due to students of average or lower ability. High-ability students—in the top 20% according to their SAT scores—were less attracted by these leniently graded courses.

## Instructors want (need) good SETs

For students' preferences of courses that promise good grades and low workloads to influence course

grades, these preferences must be adopted by faculty members and *transformed* into faculty preferences. As mentioned earlier, the medium of this transfer is the SET. Students and faculty are in an implicit negotiation situation, where each side has a "good" that is valuable to the other side. Faculty can provide good grades and easy courses, and students can provide positive SETs. Teaching quality is an important aspect of faculty evaluation and, as mentioned earlier, SETs are often used as major indicator of teaching quality. Receiving poor ratings is an unpleasant experience for teachers who are likely to have done their best to produce courses that students enjoy and in which they learn a great deal. The fact that poor SETs might decrease a teacher's chance of tenure or of a merit salary increase makes this experience even more unpleasant. Thus, the *power* of dispensing SETs—to reward desirable and to punish undesirable behavior—enables students to shape faculty behavior.

There is evidence that faculty members in precarious positions (e.g., young tenure track faculty) will be particularly motivated to improve the ratings they receive for their course by grading leniently. A study at a medium-sized state university showed a difference in the GPA of classes taught by tenured versus untenured staff of half a grade point, with untenured staff giving better grades (Moore & Trahan, 1998). Similar results were reported by Keng (2018) in a study conducted at a Taiwan university.

Faculty preference for positive SETs can influence faculty teaching by two routes—namely, a deliberate and a nondeliberate route. Surveys of instructors indicate that many teachers are aware of students' preference for leniently graded, easy courses (e.g., Birnbaum, 2000; Ryan et al., 1980; Simpson & Siguaw, 2000). For example, 65% of the faculty members who responded to a small survey conducted by Birnbaum (2000) at the California State University, Fullerton, believed that raising standards and increasing content lowered teaching evaluation, even though 45% thought that it would increase student learning. Some even admitted using strategic workload reduction and grading leniency to improve their SETs. Seventy-two percent believed that the use of SETs would encourage faculty members to water down the content of their course.

In fact, 22% of faculty respondents to a small survey conducted at the University of Wisconsin–La Crosse indicated that the introduction of SETs had induced them to decrease the amount of material covered in their course, and 38% admitted to lowering the difficulty level (Ryan et al., 1980). Finally, a web-based survey of members of the Academy of Marketing Science, which unfortunately had an extremely low response rate, asked respondents to name strategies that their colleagues had used to influence SETs. The most frequently mentioned strategy was grading leniency. Another frequently mentioned technique was serving cookies, snacks, or pizza on the day of the exam. As discussed earlier, all of these strategies are likely to be effective. If teachers are fully aware of techniques that could raise their evaluations, some of those receiving poor evaluations will be tempted to adopt these strategies, particularly if they are still untenured.

Students can also shape the behavior of faculty members in a desired direction without either side being fully aware of doing so. After receiving poor evaluations for a course, a faculty member might ask a few students what he or she should do to improve ratings. Students might complain that too much material was presented and there was too little time for discussion. They might also mention that too much reading was required for this particular course compared with other classes they were taking. The faculty member might further realize that colleagues who receive "Teacher of the Year" awards show a few films to bring theories and findings to life. To improve his or her SETs, the faculty member might therefore decide to reduce the material presented in lectures to create time for discussion, to reduce reading requirements, and to show films to make lectures more attractive. All of these changes will considerably reduce students' workload and increase the chances that they will do well in exams that will cover much less material.

It is interesting to note that, as Uttl et al. (2017) discussed, this is actually what some proponents of SETs had in mind. As Abrami and Apolonia (1990) argued,

> academic standards that are too high may be as detrimental to the learning of students as academic standards that are too low. The art of science of good teaching is finding the balance between what students might learn and what students are capable of leaning. We believe that ratings help identify those instructors who do this well. (p. 520)

So it is not the academic standard of a university or the knowledge that is required for mastering a discipline that should determine the amount that needs to be taught in a course but students' willingness to invest effort in a course and their ability to master the material. And only teachers, who are willing and able to follow these guidelines will be rewarded with good teaching evaluations.

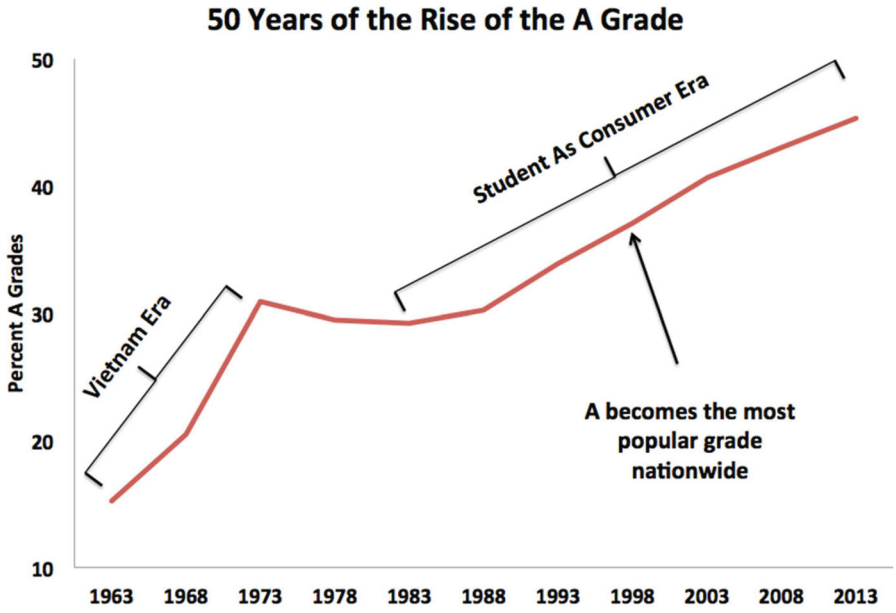


**Figure 3.** Grade inflation in the United States (from gradeinflation.com).

## Grade inflation: the evidence

In an extensive study of changes in GPAs of a large sample of private and public universities, Rojstaczer (2015; Rojstaczer & Healy, 2010) found that grades have been slowly rising since the 1930s and 1940s. However, there was a steep increase in the 1960s that leveled off in the 1970s (Figure 3). This increase has often been attributed to the Vietnam War and the wish to protect poorly performing students from being drafted (Rojstaczer & Healy, 2010). In the 1980s, the time when SETs became major information sources in faculty evaluations (Seldin, 1998), grades began to rise again at a rate of 0.10 to 0.15 GPA points per decade. These increases were much steeper for private than for public universities. Rojstaczer attributed this increase to three factors: (a) Student evaluation of classes became mandatory, (b) students became increasingly career focused, and (c) tuition rises outpaced family incomes. Students became customers and, as (paying) customers, expected a good end product—namely, a final exam with a grade that would allow them to be competitive on the job market. This interpretation would also explain why there is grade inflation in higher education in the United Kingdom (Bachan, 2017), the one European country that charges high tuition rates.

There is no evidence that students have become more intelligent or hardworking in recent decades. If anything, the evidence points in the opposite direction. Between 1969 and 1993, the average combined score on the SAT declined by 5% (Rosovsky & Harley, 2002). That this trend has continued is suggested by a report in the *Washington Post* (2015) that SAT scores had continued their downward trend. There is also no evidence that students spend more time studying. On the contrary, whereas students spent 40 hr per week on academic work in 1960, they devoted 27 hr in 2003 and 15 hr in 2008. Combining time for studying, for labs, and in classes, students today spend only 16% of their time on academic pursuits (Arum et al., 2011).

The problem with grade inflation as compared with monetary inflation is that grades are expressed on a 5-point response scale. With the upward shift in grading, it soon becomes very crowded at the top. The grade inflation was mainly driven by an increase in As. In 2006, 43% of all letter grades were As, an increase of 28 percentage points since 1960 and of 12 percentage points since 1988 (Rojstaczer & Healy, 2010). The strongest effect of grade inflation occurred for private universities, where As and Bs became even more prevalent than at public universities. At Harvard, the percentage of As for undergraduate courses increased from 22% in 1966 to 46% in 1996–1997 (B.P. Wilson, 1998). Most striking, by 2013, A– had become the median grade for undergraduates (Bernhard, 2014).

Grade inflation appears to have mainly affected the humanities and social sciences (with the exception of economics). There appears to be much less grade inflation in physics, chemistry, and mathematics. It is not clear why this is the case. As reported earlier, there certainly is no indication that students in those "hard science" areas are uninterested in receiving

good grades (Uttl & Smibert, 2017). So why do these professors not ease course loads and grade more leniently? One explanation could be that classes in hard sciences and mathematics have much more clearly defined teaching goals. Lodahl and Gordon (1972) use Kuhn's (1964) concept of paradigm to describe this difference: "The essence of a paradigm concept is the degree of consensus or sharing of beliefs within a scientific field about theory, methodology, techniques and problems" (Lodahl & Gordon, 1972, p. 58). High paradigm fields will therefore have greater agreement about the content of a course. Whereas instructors have a great deal of freedom in deciding how much a student has to learn in an introductory psychology class, this might not be the case in a class teaching mathematics, physics, or chemistry.

## The dark side of grade inflation

One could argue that grade inflation is a win-win situation. Students receive good grades and instructors receive good SETs and everybody is happy. Unfortunately, there is a dark side to grade inflation and the actual effects are much less positive. Grade inflation reduces the incentive to excel, or even to work reasonably hard. Why should a student invest a great deal of time in working for a class, if everybody gets an A? It is therefore not surprising that students have been found to work less hard in leniently graded classes (Babcock, 2010; Greenwald & Gillmore, 1997b).

For example, Greenwald and Gillmore (1997b) added questions about expected grades in a course and about workload (numbers of hours students spent on a course) to a standard SET. Whereas one would think that students would work harder if they expected a good grade in a course, this should not apply if this grade expectation is based on knowledge about grading leniency. Consistent with this latter hypothesis, Greenwald and Gillmore (1997b) found a negative correlation between workload and expected grades. This effect became even stronger when relative expected grade was used as a correlate. The better students expected to do in the present course compared with their usual performance, the less time they invested in coursework. This finding was replicated by Babcock (2010) with nearly 8,000 classes covering the years 2003 to 2007. Babcock estimated that a 1-point increase in expected grade would reduce weekly study time by 0.94 hr.

If students evaluate a course more positively the more leniently they are graded (e.g., Griffin, 2004; Olivares, 2001), but at the same time work less in such courses, one could expect that students learn less in their more positively evaluated courses. This would suggest that good teaching evaluations may, in fact, reward bad teaching, at least if one defines "bad teaching" as courses in which students do not learn a great deal (Stroebe, 2016). The fact that course grades can no longer be considered a valid indicator of learning leaves us without a measure of teaching effectiveness. Johnson (2003) suggested the brilliant but simple solution to take students' performance in follow-up courses as a measure of learning. Thus, the amount students learned in an introductory statistics course should be related to the grades they receive in an advanced statistics course that builds on the knowledge acquired in the introduction.

Six studies that used this paradigm have been conducted (Braga et al., 2014; Carrell & West, 2010; Johnson, 2003; Keng, 2018; Weinberg et al., 2009; Yunker & Yunker, 2003). Most of them could rely on large numbers of course sections. This is important, because SETs are anonymous and ratings are not known for individual students. Therefore, the association of SETs to grades in concurrent courses is typically computed as correlation of the average SET of a section with the average grades of that section. Although the study by Yunker and Yunker (2003) is based on only 46 sections taught by 12 faculty members, the more typical study of Carrell and West (2010)—conducted at the U.S. Air Force Academy—had a sample size of 10,534 students taught in 2,820 separate course sections by 421 faculty members. This study and the study by Braga et al. (2014) also have the advantage that students were randomly assigned to sections. In all six studies, SETs were positively correlated with students' grades in the concurrent course. However, when performance in subsequent—more advanced—courses was used as criterion, course ratings were negatively correlated in five of the studies (Braga et al., 2014; Carrell & West, 2010; Johnson, 2003; Keng, 2018; Yunker & Yunker, 2003).[4] In the Weinberg et al. (2009) study, student evaluations of the current course were found unrelated to the performance in a subsequent course.

In summary, then, with the exception of the study by Weinberg et al. (2009), these studies present evidence that students tend to evaluate more positively the courses in which they did not learn a great deal. Or, as Braga et al. (2014) concluded, "teachers, who are more effective in promoting future performance receive worse evaluations from their students" (p. 81).

## General conclusion

The main conclusion to be derived from the evidence reviewed in this article is that SETs are not valid measures of teaching effectiveness and should therefore not be used by deans or chairpersons to evaluate faculty members. University administrators, who still base personnel decisions on evidence from SETs, run the risk that their decisions could be challenged in court. Given the evidence that SETs are invalid as measures of teaching effectiveness—and as the example of Ryerson University suggests—the chances of winning such court cases are not very good.

Because teaching ability is an important factor in decisions about merit increases or promotions, university administrators could use alternative sources of information. Instead of using SETs, they could ask teachers to compile teaching portfolios in which they give detailed descriptions of how they develop their courses and which issues they emphasize; the portfolio should also contain lists of recommended reading and exam questions. This would at least ensure that a course on a given area covers the content (i.e., theories, research) considered central to that area and is based on up-to-date literature. In addition, such a portfolio could include the PowerPoint presentations of some of the main lectures (e.g., Goss & Bernstein, 2015). In the case of tenure decisions, one could video record a few lectures for evaluation by senior colleagues. Because the decision about teaching would only involve identifying a very poor performance, it should be possible to reach agreement in such evaluations.

SET information should be provided only to the instructors who are evaluated. This would substantially reduce the pressure on faculty members to receive top SET scores (e.g., by lowering standards and lenient grading). Although SETs are biased, totally abandoning them would deprive not only students of their voice with regard to teaching quality but also instructors of information they might find useful. However, instructors should be aware of the various sources of bias that affect SETs and take them into account when assessing the evaluation of their own teaching. There are some inspired teachers who receive top ratings even though they are tough graders, require a lot of reading, look for hard work from their students. However, for most instructors such strategies will typically result in less-than-optimal SETs, at least when they teach undergraduate courses. Yet, based on a large study of undergraduate learning at college campuses, Arum and Roksa (2011) concluded "that students, who took courses requiring both significant reading (more than 40 pages a week) and writing (more than 20 pages over the course of the semester) had higher rates of learning" (p. 205).

Because students have become used to getting good grades, at least in the social sciences and humanities, radical measures would be needed to eliminate grade inflation. The introduction of grade ceilings is one effective strategy (Butcher et al., 2014; Gorry, 2017). Unfortunately, grade ceilings have the disadvantage of lowering the competitiveness of students from universities practicing this policy. Because employers might not be aware of the Princeton grade ceiling, they might prefer a Harvard graduate with an A– to a Princeton graduate with a B+, even if the B+ might have been more difficult to get than the A–. This is probably the reason why Princeton dropped this measure in 2014 (Windemuth, 2014), and why Wellesley College dropped it in 2019. Thus, grade ceilings would be acceptable only if they were introduced by the whole system of private and public universities. In the United States, something that is unlikely to happen. Instead, a less ambitious measure that could generally be instituted and would make grades more informative for potential employers would be to indicate the median grade of a class with students' individual grades on their transcripts. It is ironic, though, that SETs intended to provide university teachers with information about how their teaching was perceived by students and thus help them to make improvements resulted not only in grade inflation but also in a deterioration of teaching quality.

## Notes

1. In May 2016, RateMyProfessors.com changed their rating scheme. They dropped the clarity and helpfulness scores and now ask students to explicitly rate the overall quality of professors. In June 2018 they also dropped the hotness ratings, responding to complaints that it was sexist.
2. The description of the study in the article is not totally clear. The authors refer to supplementary material, which I was unable to access. The address for supplementary material given in the article always brought up the article but not the material. An email to Dr. Mitchell was not answered.
3. This paradox may not be a uniquely American phenomenon. There is also evidence of grade inflation in Great Britain (Adams, 2019; Bachan, 2017; Stroebe, 2016). However, there is much less research about underlying processes.
4. Because students' grades are also strongly influenced by ability factors, Yunker and Yunker (2003) used the overall GPA as well as the ACT score as controls. Weinberg et al. (2009) controlled for grades in the current course, which is problematic.

## References


Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analyses. *Review of Educational Research*, *58*(2), 151–179. https://doi.org/10.3102/00346543058002151

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology*, *82*(2), 219–231. https://doi.org/10.1037/0022-0663.82.2.219

Adams, C. J. (2015). Latest SAT scores continue downward trend, College Board reports. *Education Week*. https://blogs.edweek.org/edweek/high_school_and_beyond/2015/09/latest_sat_scores_continue_downward_trend_college_board_reports.html

Adams, R. (2019). *Grade inflation fears prompt new voluntary code for UK degrees*. https://www.theguardian.com/education/2019/oct/10/grade-inflation-fears-prompt-new-voluntary-code-uk-degrees

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teachers evaluations from thin slices of nonverbal behavior and physical attaractiveness. *Journal of Personality and Social Psychology*, *64*(3), 431–441. https://doi.org/10.1037/0022-3514.64.3.431

Anderson, R. E., Choi, K. S., & Hair, J. F. (1975). Cognitive consistency theory and student evaluation of teacher effectiveness. *The Journal of Experimental Education*, *44*(2), 64–70. https://doi.org/10.1080/00220973.1975.11011526

Arceo-Gomez, E. O., & Campos-Vazquez, R. M. (2019). Gender stereotypes: The case of MisProfesores.com in Mexico. *Economics of Education Review*, *72*, 55–65. https://doi.org/10.1016/j.econedurev.2019.05.007

Arum, R., & Roksa, J. (2011). Limited learning on college campuses. *Society*, *48*(3), 203–207. https://doi.org/10.1007/s12115-011-9417-8

Arum, R., Roksa, J., & Cho, E. (2011). *Improving undergraduate learning: Findings and policy recommendations from the SSRC-CLA longitudinal project*. Social Science Research Council.

Babcock, P. (2010). Real costs of nominal grade inflation? New evidence from student course evaluations. *Economic Inquiry*, *48*(4), 983–996. https://doi.org/10.1111/j.1465-7295.2009.00245.x

Babcock, P., & Marks, M. (2011). The falling time cost of college: Evidence from half a century of time use data. *Review of Economics and Statistics*, *93*(2), 468–478. https://doi.org/10.1162/REST_a_00093

Bachan, R. (2017). Grade inflation in UK higher education. *Studies in Higher Education*, *42*(8), 1580–1600. https://doi.org/10.1080/03075079.2015.1019450

Bar, T., Kadiyali, V., & Zussman, A. (2009). Grade information and grade inflation: The Cornell experiment. *Journal of Economic Perspectives*, *23*(3), 93–108. https://doi.org/10.1257/jep.23.3.93

Bekelman, J. E., Li, Y., & Gross, C. P. (2003). Scope and impact of financial conflicts of interest in biomedical research: A systematic review. *JAMA*, *289*(4), 454–465. https://doi.org/10.1001/jama.289.4.454

Bernhard, M. P. (2014). Princeton grade deflation reversal disappoints some here. *Harvard Crimson*. http://www.thecrimson.com/article/2014/10/9/princeton-grade-deflation-reversal.

Birnbaum, M. (2000). *A survey of faculty opinions concerning student evaluations of teaching*. http://psych.fullerton.edu/mbirnbaum/faculty3.htm.

Boehmer, D. M., & Wood, W. C. (2017). Student vs. faculty perspectives on quality instruction: Gender bias, "hotness", and "easiness" in evaluating teaching. *Journal of Education for Business*, *92*(4), 173–178. https://doi.org/10.1080/08832323.2017.1313189

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, *145*, 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *Science Open Research*. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, *41*, 71–88. https://doi.org/10.1016/j.econedurev.2014.04.002

Brown, M. J., Baillie, M., & Fraser, S. (2009). Rating RateMyProfessors.com: A comparison of online and official student evaluations of teaching. *College Teaching*, *57*(2), 89–92. https://doi.org/10.3200/CTCH.57.2.89-92

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. https://doi.org/10.1257/jep.28.3.189

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.

Centra, J. A. (2009). *Differences in responses to the student instructional report: Is it bias?* Educational Testing Service.

Chiu, Y.-L., Chen, K.-H., Hsu, Y.-T., & Wang, J.-N. (2019). Understanding the perceived quality of professors' teaching effectiveness invarious disciplines: The moderating effects of teaching at top colleges. *Assessment & Evaluation in Higher Education*, *44*(3), 449–462. https://doi.org/10.1080/02602938.2018.1520193

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn?. *Journal of Marketing Education*, *31*(1), 16–30. https://doi.org/10.1177/0273475308324086

Clayson, D. E., Frost, T. F., & Sheffet, M. J. (2006). Grades and the student evaluation of instruction: A test of the reciprocity effect. *Academy of Management Learning & Education*, *5*(1), 52–65. https://doi.org/10.5465/amle.2006.20388384

Cohen, P. A. (1981). Student ratings of instruction and student achievements: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51*(3), 281–309. https://doi.org/10.3102/00346543051003281

Cohen, P. A. (1983). Comment on a selective review of the validity of student ratings of *teaching*. *Journal of Higher Education*, *54*, 78–82.

Colardarci, T., & Kornfield, I. (2007). RateMy Professors.com versus formal, in-class student evaluations of teaching. *Practical Assessment, Research and Evaluation*, *44*(12), 1–15.

Delucchi, M. (2000). Don't worry. Be happy: Instructor likability, student perceptions of learning, and teacher ratings in upper-level sociology courses. *Teaching Sociology*, *28*(3), 220–231. https://doi.org/10.2307/1318991

Eagly, A., Ashmore, R. D., Makhijani, M. G., Longo, L. C. (1991). What is beautiful is good, but … : A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*(1), 109–128. https://doi.org/10.1037/0033-2909.110.1.109

Feistauer, D., & Richter, T. (2018). Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest. *Studies in Educational Evaluation*, *59*, 168–178. https://doi.org/10.1016/j.stueduc.2018.07.009

Feldman, K. A. (1989). The association between student rating of specific instructional dimensional student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, *30*(6), 583–645. https://doi.org/10.1007/BF00992392

Felton, J., Koper, P. T., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on Ratemyprofesssors.com. *Assessment & Evaluation in Higher Education*, *33*(1), 45–61. https://doi.org/10.1080/02602930601122803

Fisher, A. N., Stinson, D. A., & Kalajdzic, A. (2019). Unpacking backlash: Individual and contextual moderators of bias against female professors. *Basic and Applied Social Psychology*, *41*(5), 305–325. https://doi.org/10.1080/01973533.2019.1652178

Freishtat, R. L. (2016). *Expert report on student evaluations of teaching (SET)*. https://ocufa.on.ca/assets/RFA.v.Ryerson_Stark.Expert.Report.2016.pdf?utm_source=OCUFA+Report&utm_campaign=7bb120ce70-EMAIL_CAMPAIGN_2018_07_12_01_15&utm_medium=email&utm_term=0_458512323c-7bb120ce70-&mc_cid=7bb120ce70&mc_eid=%5BUNIQID%5D

Freng, S., & Webber, D. (2009). Turning up the heat on online teaching evaluations: Does "hotness" matter. *Teaching of Psychology*, *36*(3), 189–193. https://doi.org/10.1080/00986280902959739

Gorry, D. (2017). The impact of grade ceilings on student grades and course evaluations: Evidence from a policy change. *Economics of Education Review*, *56*, 133–140. https://doi.org/10.1016/j.econedurev.2016.12.006

Goss, L. S., & Bernstein, D. A. (2015). *Teaching psychology: A step by step guide* (2nd ed.). Taylor & Francis.

Greenwald, A. G., & Gillmore, G. M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *52*(11), 1209–1217. https://doi.org/10.1037/0003-066X.52.11.1209

Greenwald, A. G., & Gillmore, G. M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instructions. *Journal of Educational Psychology*, *89*(4), 743–751. https://doi.org/10.1037/0022-0663.89.4.743

Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instructions. *Contemporary Educational Psychology*, *29*(4), 410–425. https://doi.org/10.1016/j.cedpsych.2003.11.001

Gurung, R. A. R., & Vespia, K. M. (2007). Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology*, *34*(1), 5–10. https://doi.org/10.1207/s15328023top3401_2

Guthrie, E. R. (1953). The evaluation of teaching. *The American Journal of Nursing*, *53*(2), 220–221. https://doi.org/10.2307/3459921

Hamermesh, D. S., & Parker, A. (2005). Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review*, *24*(4), 369–376. https://doi.org/10.1016/j.econedurev.2004.07.013

Hessler, M., Pöpping, D. M., Hollstein, H., Ohlenburg, H., Arnemann, P. H., Massoth, C., Seidel, L. M., Zarbock, A., & Wenk, M. (2018). Availability of cookies during an academic course session affects evaluation of teaching. *Medical Education*, *52*(10), 1064–1072. https://doi.org/10.1111/medu.13627

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

Johnson, R. R., & Crews, A. D. (2013). My professor is hot! Correlates of RateMyProfessors.com ratings for criminal justice and criminology faculty members. *American Journal of Criminal Justice*, *38*(4), 639–656. https://doi.org/10.1007/s12103-012-9186-y

Keng, S. H. (2018). Tenure system and its impact on grading leniency, teaching effectiveness and student effort. *Empirical Economics*, *55*(3), 1207–1227. https://doi.org/10.1007/s00181-017-1313-7

Legg, A. M., & Wilson, J. H. (2012). RateMyProfessors.com offers biased evaluations. *Assessment & Evaluation in Higher Education*, *37*(1), 89–97. https://doi.org/10.1080/02602938.2010.507299

Lodahl, J. B., & Gordon, G. (1972). The structure of scientific fields and the functioning of university graduate departments. *American Sociological Review*, *37*(1), 57–72. https://doi.org/10.2307/2093493

Lundh, A., Lexchin, J., Mintzes, B., Schroll, J. B., & Bero, L. (2018). Industry sponsorship and research outcome: A systematic review with meta-analysis. *Intensive Care Medicine*, *44*(10), 1603–1612. https://doi.org/10.1007/s00134-018-5293-7

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in Student Rating of Teaching. *Innovative Higher Education*, *40*(4), 291–303. https://doi.org/10.1007/s10755-014-9313-4

McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education*, *21*(2), 150–158. https://doi.org/10.1007/BF00975102

McPherson, M. A., & Jewell, R. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly*, *88*(3), 868–881. https://doi.org/10.1111/j.1540-6237.2007.00487.x

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluation. *Journal of the European Economic Association*, *17*(2), 535–566. https://doi.org/10.1093/jeea/jvx057

Miller, J. E., Seldin, P. (2014). *Changing Practices in Faculty Evaluation: Can better evaluation make a difference?* American Association of University Professors. http://www.aaup.org/article/changing-practices-faculty-evaluation#.VulYjE0UWpo

Mitchell, K. M. W., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, *51*(3), 648–652. https://doi.org/10.1017/S104909651800001X

Moore, M., & Trahan, R. (1998). Tenure status and grading practices. *Sociological Perspectives*, *41*(4), 775–781. https://doi.org/10.2307/1389669

Murray, K. B., & Zdravkovic, S. (2016). Does MTV really do a good job of evaluating professors? An empirical test of the internet site RateMyProfessors.com. *Journal of Education for Business*, *91*(3), 138–147. https://doi.org/10.1080/08832323.2016.1140115

Olivares, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology*, *26*(3), 382–399. https://doi.org/10.1006/ceps.2000.1070

Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review*, *29*(6), 923–934. https://doi.org/10.1016/j.econedurev.2010.06.011

Pascarella, E. T., Blaich, C., Martin, G. L., & Hanson, J. M. (2011). How robust are the findings of Academically Adrift? *Change: The Magazine of Higher Learning*, *43*(3), 20–24. https://doi.org/10.1080/00091383.2011.568898

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors. *Journal of Diversity in Higher Education*, *3*(3), 137–153. https://doi.org/10.1037/a0019865

Remmers, H. H., & Brandenburg, G. C. (1927). Experimental data on the Purdue Rating Scale for Instruction. *Educational Administration and Supervision*, *13*, 519–527.

Reysen, S. (2005). Construction of a new scale: The Reysen likability scale. *Social Behavior and Personality: An International Journal*, *23*, 2001–2208.

Riniolo, T. K., Johnson, k., Sherman, T., & Misso, J. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations. *The Journal of General Psychology*, *133*(1), 19–35. https://doi.org/10.3200/GENP.133.1.19-35

Rojstaczer, S. (2015). *Grade inflation at American colleges and universities*. http://:www.gradeinflation.com

Rojstaczer, S., Healy, C. (2010). *Where a is ordinary: The evolution of american college and university grading, 1940–2009* (Teachers College Record, Date Published: March 4, 2010). http://www.tcrecord.org.

Rosen, A. S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data. *Assessment & Evaluation in Higher Education*, *43*(1), 31–44. https://doi.org/10.1080/02602938.2016.1276155

Rosovsky, H., Harley, M. (2002). *Evaluation and the Academy: Are we doing the right thing?* Academy of Arts and Sciences. https://www.amacad.org/multimedia/pdfs/publications/researchpapersmonographs/Evaluation_and_the_Academy.pdf.

Ryan, J. J., Anderson, J. A., & Birchler, A. B. (1980). Student evaluation: the faculty responds. *Research in Higher Education*, *12*(4), 317–333. https://doi.org/10.1007/BF00976185

Ryerson University v Ryerson Faculty Association. (2018). https://www.psychologicalscience.org/redesign/wp-content/uploads/2019/05/MayJune_OBS_2019-OnlineSmall.pdf

Sabot, R., & Wakeman-Linn, J. (1991). Grade inflation and course choice. *Journal of Economic Perspectives*, *5*(1), 159–170. https://doi.org/10.1257/jep.5.1.159

Samuelson, P., A. (1948). Consumption theory in terms of revealed preference. *Economica*, *15*(60), 243–253. https://doi.org/10.2307/2549561.JSTOR

Seldin, P. (1998). How colleges evaluate teaching: 1988 vs. 1998. *American Association of Higher Education Bulletin*, *50*, 3–7.

Simpson, P., & Siguaw, J. A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education*, *22*(3), 199–213. https://doi.org/10.1177/0273475300223004

Sinclair, U. (1994). *I, candidate for governor: And how I got licked*. University of California Press. (Original work published 1934)

Sismondo, S. (2008). How pharmaceutical industry funding affects trial outcomes: Causal structure and responses. *Social Science and Medicine*, *353*, 1060–1065.

Smith, B. P. (2007). Student ratings of teacher effectiveness: An analysis of end-of-curse faculty evaluation. *College Student Journal*, *41*, 788–800.

Sonntag, M. E., Bassett, J. R., & Snyder, T. (2009). An empirical test of the validity of student evaluations of teaching made on RateMyProfessors.com. *Assessment & Evaluation in Higher Education*, *34*(5), 499–504. https://doi.org/10.1080/02602930802079463

Stuber, J. M., Watson, A., Carle, A., & Staggs, K. (2009). Gender expectation and on-line evaluation of teaching: Evidence from RateMyProfessors.com. *Teaching in Higher Education*, *14*(4), 387–399.

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science*, *11*(6), 800–816. https://doi.org/10.1177/1745691616650284

Timmerman, T. (2008). On the validity of RateMyProfesors.com. *Journal of Education for Business*, *84*(1), 55–61. https://doi.org/10.3200/JOEB.84.1.55-61

Uttl, B., Cnudde, K., & White, C. M. (2019). Conflict of interest explains the size of student evaluation of teaching and learning correlations in multisection studies: a meta-analysis. *PeerJ*, *7*, e7225. https://doi.org/10.7717/peerj.7225

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. *Peer J*, *5*, e3299.

Uttl, B., & Violo, V. C. (2020). Small samples, unreasonable generalizations, and outiers: Gender bias in student evaluation of teaching or three unhappy students? *ScienceOpen Preprints*. https://doi.org/10.14293/S2199-1006.1.SOR-.PPUTIGR.v1

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching rating and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42.

Vartanian, L., Schwartz, M. B., & Brownell, K. D. (2007). Effects of soft drink consumption on nutrition and health: A systematic review and meta-analysis. *American*

*Journal of Public Health*, *97*(4), 667–675. https://doi.org/10.2105/AJPH.2005.083782

Washington Post. (2015). *SAT scores at lowest level in 10 years, fueling worries about high school*. https://www.washingtonpost.com/local/education/sat-scores-at-lowest-level-in-10-years-fueling-worries-about-high-schools/2015/09/02/6b73ec66-5190-11e5-9812-92d5948a40f8_story.html?utm_term=.c5f0ac7caf02

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *The Journal of Economic Education*, *40*(3), 227–261. https://doi.org/10.3200/JECE.40.3.227-261

Wilson, B. P. (1998). *The phenomenon of grade inflation in higher education*. http://www.virginiaeducators.org/gradeinflation.html

Windemuth, A. (2014). After faculty vote, grade deflation policy officially dead. *Daily Princetonian*. http://dailyprincetonian.com/news/2014/10/breaking-after-faculty-vote-grade-deflation-policy-officially-dead/

Wolbring, T., & Riordan, P. (2016). How beauty works. Theoretical mechanism and two empirical applications on students' evaluation of teaching. *Social Science Research*, *57*, 253–273. https://doi.org/10.1016/j.ssresearch.2015.12.009

Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course. *Teaching of Psychology*, *34*(4), 245–247. https://doi.org/10.1080/009862807 01700318

Yunker, P. J., & Yunker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business*, *78*(6), 313–317. https://doi.org/10.1080/08832320309598619