aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2016, Vol. 11(6) 800–816

DOI: 10.1177/1745691616650284
pps.sagepub.com
SAGE

Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations

Wolfgang Stroebe
Department of Social and Organizational Psychology, University of Groningen, the Netherlands

Abstract

In this article, I address the paradox that university grade point averages have increased for decades, whereas the time students invest in their studies has decreased. I argue that one major contributor to this paradox is grading leniency, encouraged by the practice of university administrators to base important personnel decisions on student evaluations of teaching. Grading leniency creates strong incentives for instructors to teach in ways that would result in good student evaluations. Because many instructors believe that the average student prefers courses that are entertaining, require little work, and result in high grades, they feel under pressure to conform to those expectations. Evidence is presented that the positive association between student grades and their evaluation of teaching reflects a bias rather than teaching effectiveness. If good teaching evaluations reflected improved student learning due to effective teaching, they should be positively related to the grades received in subsequent courses that build on knowledge gained in the previous course. Findings that teaching evaluations of concurrent courses, though positively correlated with concurrent grades, are negatively related to student performance in subsequent courses are more consistent with the assumption that concurrent evaluations are the result of lenient grading rather than effective teaching. Policy implications are discussed.

Keywords

bias, grade inflation, grade point average, grading leniency, study time investment, teacher effectiveness

It is one of the great paradoxes of higher education in the United States that the grade point average (GPA) at colleges and universities has increased for decades (e.g., Rojstaczer, 2015; Rojstaczer & Healy, 2010), whereas the amount of time students devote to their studies has continuously decreased (Arum & Roksa, 2011; Babcock & Marks, 2011; Pascarella, Blaich, Martin, & Hanson, 2011). This increase in the GPA of students over an extended period of time without a corresponding increase in student achievement has been referred to as *grade inflation* (e.g., Rojstaczer, 2015; Rojstaczer & Healy, 2010; Rosovsky & Hartley, 2002). From their analysis of changes of GPA from 1930 to 2006 in a large sample of public and private universities across the United States, Rojstaczer (2015) concluded that grades were rising already in the 1930s and 1940s. However, there was a steep increase in the 1960s that leveled off in the 1970s. Grades began to rise again in the 1980s and increased at a rate of about 0.10 to 0.15 GPA points per decade. These increases were steeper for private than for public colleges (Table 1; Rojstaczer, 2015). The 1980s are also the period when the use of student evaluations of teaching as a measure of faculty evaluation became standard practice (Seldin, 1998). Similar changes in GPA were reported by Jewell, McPherson, and Tieslau (2013) based on a study at the University of North Texas that covered all courses over a period of 21 academic years (1984–1985 to 2004–2005). GPA increased from 2.49 in 1984 to 2.86 in 2005. The problem with grade inflation compared with monetary inflation is the limited response scale used in grading.

Corresponding Author:
Wolfgang Stroebe, Department of Social and Organizational Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712TS Groningen, the Netherlands
E-mail: W.Stroebe@uu.nl

Table 1. Historical Increase in GPA at American Colleges and Universities (Adapted From Rojstaczer, 2015)

Academic year	1991–1992	1996–1997	2001–2002	2006–2007
All schools	2.93	2.99	3.07	3.11
Public schools	2.85	2.90	2.97	3.01
Private schools	3.09	3.16	3.24	3.30

With more and more students receiving top grades, grades lose their ability to distinguish the excellent from the merely good or even the only moderate students.

The grade inflation seems even more substantial if one considers changes in letter grades. By 2006, 43% of all letter grades were As. This represents an increase of 28 percentage points since 1960 and 12 percentage points since 1988 (Rojstaczer, 2015; Rojstaczer & Healy, 2010). Furthermore, private colleges and universities gave significantly more As and Bs than public institutions with equal student selectivity (Rojstaczer & Healy, 2010). At Harvard, the percentage of A and A– undergraduate grades increased from 22% in 1966 to 46% during the 1996–1997 year (B. P. Wilson, 1998). By 2013, A– had become the median grade for undergraduates at Harvard (Bernhard, 2014).

Rosovsky and Hartley (2002) argued that unless there had been an extraordinary improvement in the quality of students during this period, the increase in GPA has to be interpreted as inflation. In addition, there is no indication for such an improvement. In fact, from 1969 to 1993, the average combined score on the Scholastic Achievement Test (SAT) declined by 5% (Rosovsky & Hartley, 2002), and there is no evidence of major improvements since then (Jacobsen, 2015). There is also no evidence that students increased the time they invest in their studies. On the contrary, the time full-time college students allocated toward class and studying decreased from 40 hr per week in 1961 to 27 hr in 2003 (Babcock & Marks, 2011). More recent surveys even suggest further reduction. On the basis of a longitudinal study of more than 3,000 students from Fall 2005 to Spring 2009 across 29 four-year colleges and universities, Arum and Roksa (2011) concluded that "on average, students in a typical semester spend only between 12 and 14 h per week studying . . . Combining the hours spent studying with the hours spent in classes and labs, students spend less than one-fifth (16%) of their time each week on academic pursuits" (p. 204). On the basis of a similar study, Pascarella et al. (2011) concluded that the "typical full-time student spent about 15 hours per week preparing for class" (p. 23). Consistent with this, there appears to have been a considerable reduction in students' 4-year gains in thinking skills during the last few decades (Arum & Roksa, 2011; Pascarella et al., 2011).

In this article, I argue that the widespread practice of giving great weight to student evaluations of teaching (SET) in decisions about salary increases, promotion, tenure, and even appointments of new faculty members is one of the major reasons for both grade inflation and the decrease in the time students invest in their studies. When SETs became important determinants of academic personnel decisions, strong incentives were created for faculty to teach in ways that would result in good student evaluations. Since many faculty members believe that the average student prefers courses that require little work and result in high grades (Birnbaum, 2000; Ryan, Anderson, & Birchler, 1980; Simpson & Siguaw, 2000), one way to improve ratings was to grade more leniently and lower the demands for work to be done for a course. This argument has been made by others (e.g., Greenwald & Gillmore, 1997a, 1997b; Johnson, 2003; Krautmann & Sander, 1999) but has never been fully developed theoretically. Furthermore, a great deal of new evidence has emerged that supports this hypothesis (e.g., Babcock, 2010; Braga, Paccagnella, & Pellizzari, 2014; Carrell & West, 2010; Felton, Koper, Mitchell, & Stinson, 2008; Johnson, 2003; Weinberg, Hashimoto, & Fleisher, 2009; Yunker & Yunker, 2003).

I am presenting a social psychological model that integrates the different stages of this influence process. It elaborates theoretically two basic assumptions, the *bias* assumption and the *grading leniency* assumption. According to the bias assumption, the work students are required to invest in a course and the grades they receive biases their evaluation of course and instructor. The less work students have to do and the better the grade they receive, the more positive their teaching evaluation. There are two potential mechanisms for this bias to result in grade inflation: (a) Instructors are aware of the bias and use it deliberately to improve the evaluation of their teaching by lowering their course requirements and by grading more leniently. (b) Instructors who receive low teaching ratings might introduce changes, and they will retain those changes if ratings improve. Reductions in workload could be an incidental (rather than deliberate) consequence (Greenwald & Gillmore, 1997a, 1997b). For example, instructors may decide (perhaps based on student complaints) that more time needs to be spent on difficult topics. They then find that increased class time

on those difficult topics yields higher ratings (concomitantly, higher grades). In the process a few topics may be squeezed out of the syllabus, but the intent was never to decrease workload to increase ratings.

Overview

In the first section, I describe the increase in the practice of using student teaching evaluations for important academic personnel decisions and review evidence for a positive association between teaching evaluations and course grades. Because there are two explanations for this association, namely that it is because of bias and/or it is an indication of the validity of SET (i.e., reflects teaching effectiveness), I review theoretical justifications for both positions in the second section. I then present empirical research on both the bias (section 3) and the teacher effectiveness explanation (section 4). I review a great deal of recent evidence that is inconsistent with the teaching effectiveness interpretation. Particularly inconsistent are findings that even though teaching ratings are positively related to performance in the course being evaluated, they are either unrelated or negatively related to students' performance in subsequent courses that are assumed to build on knowledge acquired previously (Braga et al., 2014; Carrell & West, 2010; Johnson, 2003; Weinberg et al., 2009; Yunker & Yunker, 2003). Section 5 focuses on grading leniency. I review evidence that most teachers are aware that grades bias student evaluations and some use grading leniency to improve their teaching rating. Research further demonstrates that perceived grading leniency influences students' course ratings and can guide their course selection. In concluding sections, policy implications will be discussed.

1. A Brief History of Student Evaluations of Teaching

Studies of student evaluations of teaching were first conducted in the 1920s by the educational psychologist Hermann H. Remmers (e.g., Remmers & Brandenburg, 1927) at Purdue University and by the learning psychologist Edwin R. Guthrie (e.g., Guthrie, 1953) at the University of Washington. Originally, these student evaluations were only used to inform teachers about how their teaching was perceived by students. However, they soon became a valued source of information for university administrators. It became practice for administrators to demand "objective evidence of merit" in promotion decisions, and ratings by students became one source of evidence (Guthrie, 1953). The practice of American colleges and universities of collecting SETs increased from 29% in 1973 to 68% in 1983 and 86% in 1993 (Seldin, 1998). In addition, despite Guthrie's (1953) warning that "it would be a serious misuse of this information to accept it as an ultimate measure of merit" (p. 221), a survey of deans of private liberal arts colleges found that SETs have become the prime source of information in the evaluation of teaching and are given more weight than classroom visits or examination scores (Seldin, 1998). A 2010 survey of deans of liberal arts colleges indicated that reliance on student ratings had further increased from 88.1% to 94.2% over the 10-year period (Miller & Seldin, 2014), and 99.3% of these deans named classroom teaching as a major factor in evaluating faculty performance.

Obviously, the most critical question about student ratings of instructors is their validity as measures of teaching effectiveness. Because teaching effectiveness is generally considered in terms of the amount students learn in a course (Cohen, 1981) and course grades are the accepted measure of course learning, researchers soon began to study the association between course grades and SET as a test of the validity of student ratings. The first large-scale observational study of this association was probably D. L. Brown's (1976) doctoral dissertation at the University of Connecticut. The study was based on 2,360 course sections that were evaluated during the spring semester of 1973, providing 30,000 rating forms. The rating forms consisted of eight scales requiring evaluation of the instructor's knowledge of subject, presentation of material, balance of breadth and detail, enthusiasm for subject, fairness in marking, attitude toward student, personal mannerisms, and an overall evaluation of his or her quality as a teacher. A principal component analysis indicated that the eight items loaded on one common factor allowing for the mean of the eight rating scales being used as the criterion variable. A stepwise regression analysis using predictors such as course level, department quantitativeness, class size, or number of years instructors were tenured, accounted for 6% of the criterion variance. In a second stepwise regression, the average of the student grades in each course was added as a predictor variable. This addition significantly improved the multiple correlation from .25 to .39 and accounted for 8.5% of the variance in teaching ratings. Since then, numerous studies relating course grades to student evaluations of teaching have been conducted, and several meta-analyses have statistically combined the results of these studies (Clayson, 2009; Cohen, 1981, 1982, 1983; Dowell & Neal, 1982; McCallum, 1984). On the basis of a meta-analysis of these meta-analyses, Wright and Jenkins-Guarnieri (2012) concluded that course grades explain approximately 10% of the variance in students' teaching evaluations.

These meta-analyses combined studies that used two types of designs, computing correlations either within or between classes. Although Johnson (2003)—in a review that also included more recent studies—concluded that both types of design result in similar associations between

grades and teaching evaluation (.21 for within-class and .31 for between-class association), the distinction between the two types of design is important for theoretical reasons. With the first design, the student is the unit of analysis, and correlations are computed between individual grades and individual teaching evaluations. These studies address whether students, who receive higher grades, rate their teacher more positively. Although SETs are typically administered before students receive their final grade, there will have been one or two interim exams that give students a good idea of the grading style of their instructor as well as the overall grade they are likely to receive. Because all students in a given class are exposed to the same teaching, a positive correlation between grades and teaching rating cannot be due to differences in teaching effectiveness. With the second design, the class is the unit of analysis, and correlations are computed between class mean grades and class mean teaching evaluations. Because students in different classes are exposed to different teachers, a positive correlation between mean teaching ratings and mean course grades could reflect differences in teaching effectiveness.

Despite hundreds of studies devoted to this issue (for reviews, see e.g., Greenwald, 1997; Spooren, Brockx, & Mortelmans, 2013), the validity of student ratings of teaching as a measure of teaching effectiveness is still hotly debated. A recent review of research on the validity of SET published between the years 2000 and 2010 and based on 160 relevant publications concluded that "SET remains a current yet delicate topic in higher education, as well as in education research. Many stakeholders are not convinced of the usefulness and validity of SET for both formative and summative purposes. Research on SET has thus far failed to provide clear answers to several critical questions concerning the validity of SET" (Spooren et al., 2013, p. 598).

2. Theories of the Association of Course Grades and Teaching Evaluation

There are two potential explanations for the positive association between grades and teaching evaluation, namely that it is an indication of the validity of SET and reflects teaching effectiveness or it is the result of bias. According to a definition proposed by Marsh (1984), "student ratings are biased to the extent that they are influenced by variables unrelated to teaching effectiveness" (p. 733). Education researchers believe in the validity of student evaluations as an indication of teaching effectiveness (e.g., Centra, 2003; Kulik, 2001; Marsh, 1984, 1987; Marsh & Roche, 2000; McKeachie, 1997; K. L. Wilson, Lizzo, & Ramsden, 1997). In contrast, several statisticians, psychologists, and economists have argued that these evaluations are subject to numerous biases (e.g., Braga et al., 2014; D. L. Brown, 1976; Carrell & West, 2020; Greenwald & Gillmore, 1997a, 1997b; Johnson, 2003; Weinberg et al., 2009). Because it is possible that student evaluations of teaching are biased, even though they mainly reflect the qualities of teaching they are supposed to assess, the two explanations are not mutually exclusive. The following section reviews theories of bias as well as of teacher effectiveness.

Theories of bias

Balance theory. According to balance theory (Heider, 1958), attitudes toward persons and impersonal events are in balance, if persons and impersonal events that are perceived as belonging together are evaluated similarly. If a balanced state does not exist, there will be a tendency to restore balance. In the case of a student *P*, a class teacher *O*, and the student's test performance *X*, we are dealing with a *P*–*O*–*X* triad. In such a triad, a balanced state exists if all three relations are positive in all respects or if two are negative and one is positive (Heider, 1958). The student *P* is linked through a positive causal link to the test results he or she "caused." If the teacher *O* evaluates these results positively, then the triad is in balance, if *P* evaluates *O* positively (all links positive). However, if *O* gives a poor grade to *P*'s test, the system would be in balance, if *P* evaluates *O* poorly (i.e., two links negative, one positive). Thus, balance theory predicts a positive association between grades and teacher evaluation. Balance effects should result in a shift of evaluations away from the mean toward both the positive and negative end of the scale but leave mean ratings of teaching unaffected.

Attribution theory. According to Heider's (1958) attribution theory, actors can attribute events either to external or internal causes. Weiner (1979, 1986) extended the list of causes people use in their inferences about reasons for their success or failure, suggesting that causes can also be classified as stable or variable (stability) and controllable or uncontrollable (controllability). These three classes of perceived causes are assumed to constitute a 2 × 2 × 2 classification. Thus, if students fail in one exam, they can attribute their failure to lack of effort (internal, variable, and controllable) or bad luck (external, variable, and uncontrollable). However, if they do poorly in all examinations in a course, the only internal attribution that would still be plausible would be lack of ability (internal, stable, and uncontrollable). Because people are usually averse to thinking of themselves as stupid (e.g., Kruger & Dunning, 1999), they might search for external causes such as exams that were too difficult, tough grading, and poor teaching. Weiner's theory would predict

that students who achieve good grades attribute them to their own ability and hard work, whereas students who get poor grades would tend to blame the teacher. Attribution and balance theory are insofar related, that actors are most likely to engage in extensive search for meaning, when a situation is imbalanced and there is a need to restore balance. Attribution theory would predict a positive correlation between grades and teaching evaluation. However, because bias is only expected for poor students, this theory would predict a downward shift in mean value of the teaching evaluation. Furthermore, the distribution of teaching ratings should be skewed toward the negative endpoint of the scale.

Revenge. Unlike biases, which influence judgments without individual awareness, revenge is a conscious strategy of students aimed at punishing a teacher for poor grades. Because students are aware that SETs are used in the evaluation of teachers by department heads or deans, some particularly angry students might give poor ratings in the hope that it will have negative consequences for their instructor. According to Weiner (1986), anger is the emotional response to an attribution of blame, and one way to vent anger is to attempt revenge. The effect on ratings should be similar to that of attribution processes. However, in addition to giving poor ratings, students with a revenge motive typically also use the comments section to make negative remarks about their instructor. Because these comments are anonymous, they are a safe and easy way to punish an instructor for assigning a poor grade.

Reciprocity. Reciprocity is the opposite of the revenge motive. Students who receive good grades are assumed to be motivated to reward their teachers by giving particularly positive teaching ratings. Although the reciprocity principle derives from a different theoretical tradition (Gouldner, 1960) than balance theory (Heider, 1958) and does not presume liking as a precondition for reciprocity, the predictions for bias in teaching evaluation are similar. According to the norm of reciprocity, "when one party benefits another, an obligation is generated. The recipient is now indebted to the donor, and he remains so until he repays" (Gouldner, 1960, p. 171). Because grades are not gifts but reflect the instructor's ratings of the quality of a student's performance, one can doubt whether a student, who receives a good grade—even if undeserved—feels obliged to reward the instructor with a good teaching evaluation. Furthermore, because teaching ratings are anonymous, the donor would never realize that an obligation has been repaid. I would therefore argue that reciprocity does not constitute an appropriate explanation for the positive association of course grades and student ratings.

A theoretical integration. When students who think of themselves as competent receive good grades, the situation is balanced if they perceive the instructor as a competent and good teacher. They will attribute the grade internally and feel no need to analyze the situation further. Any attractive features of the instructor (e.g., physical attractiveness, sense of humor) will increase their liking. In contrast, if a teacher assigns poor grades, the system is imbalanced, and students will feel the need to interpret the causes of the negative event. One way to achieve balance is to make external attributions and blame the poor test results on the instructor, who taught them badly. Because they perceive their grades as unjust, they might be angry (Weiner, 1986) and even feel the need for revenge. However, not all students are likely to react this way. The findings of Kruger and Dunning (1999) suggest that this way of responding is particularly characteristic of less competent students. Good students are more likely to acknowledge that a poor test result could have been avoided had they worked harder for a course.

Teacher effectiveness theory

According to the teacher effectiveness theory, the positive association between teacher ratings and grades demonstrates the validity of SET and is due to the fact that students learn more when taught well and therefore receive better grades in classes taught by effective teachers (Johnson, 2003). The teacher effectiveness theory predicts a between-class correlation of course mean grades and course mean ratings of teaching. Because all students in a class are exposed to the same teacher, there should be no within-class correlation of individual grades and teaching evaluation.

3. Bias in Teaching Evaluation: Empirical Evidence

In this section, I will first review empirical evidence from two types of sources (experiments, correlational evidence) that offer some support for bias. I will then discuss tests of some of the theoretical interpretations of the bias.

Experimental studies

The best way to examine whether there is a causal influence of grades on teaching evaluation would be experimental studies that manipulate the grades students receive and assess the effect of this manipulation on teaching evaluation. Realistic field experiments that use this type of false feedback would never pass an ethics board today but could still be done in the 1970s, when

most departments had not yet instituted ethics boards. In what is probably the first experiment of this type, Holmes (1972) studied the effect of disconfirmed grade expectations on teaching evaluation. Students in an introductory psychology class were given final course grades, which were either the grade they expected or a grade that was one step below their expectations. Grades in this course were based on four multiple-choice exams spread throughout the semester. After the third exam, students' grade expectations were measured. After the fourth exam, students who deserved and expected As or Bs and were either given their expected grade or a lower grade responded to a 19-item SET. A 2 (actual A vs. B) by 2 (expected grade vs. lower grade) analysis of variance (ANOVA) revealed no effect of actual grades but significant effects of expected grades on some of the rating scales: Students whose grade expectations had been disappointed gave the instructor significantly poorer evaluations on 5 out of 19 SET items and marginally significant ones on five more items. Unfortunately Holmes (1972) did not conduct multivariate ANOVAs to examine the overall significance of his manipulation. That actual grades had no effect could have been because it could only be tested for As and Bs. Holmes thought that students getting Cs could not be credibly manipulated to the D level.

A study by Worthington and Wong (1979) attempted to extend the Holmes study by not only lowering but also increasing student grades. In this study, students were given a short multiple choice test based on the lecture content and reading for the material covered in the lecture section of the course. Students were informed that this test would not count toward their final grades but would serve as an indication of how well they were likely to do. Students were divided into three groups based on their performance in the practice test representing the top, middle, and lower segments of the distribution. Members of each of these segments were then randomly assigned a grade of good, satisfactory, or poor. Although the design called for a factorial multivariate analysis of variance, the authors decided to conduct numerous contrasts instead, because absenteeism had resulted in unequal cell sizes (which were difficult to handle statistically in those days). Because they also conducted independent tests with each of the contrasts for each of the 23 ratings scales that constituted their SET, one cannot trust the few significant results that occurred.

In another study of the effect of experimental grade manipulation on teacher ratings, Vasta and Sarmiento (1979) manipulated the strictness of grading in two large sections of an undergraduate psychology course. Students received four multiple choice tests throughout the semester, and although the average numerical grade of the two sections did not differ significantly, the letter grades were manipulated by imposing either more liberal or stricter grading norms. Although Vasta and Sarmiento again did independent statistical tests on the items that evaluated the course and the instructors, they also conducted nonparametric checks on the probability that 16 of the 18 items evaluating the course and 24 of the 32 items evaluating the instructor were more positive in the section that had received the more lenient grading. In both cases they found that the probability of receiving such a distribution by chance was less than .01.

The most persuasive finding comes from a study that was not a proper experiment but made use of an anti-grade–inflation policy instituted at Wellesley College (Butcher, McEwan, & Weerapana, 2014). In the early 2000s, the faculty and administration at Wellesley concluded that grade inflation was undermining the institution's credibility. As was generally the case, grade inflation did not affect science departments but was a problem in most other departments. The college therefore instituted the rule that average grades must not be higher than 3.33 (B+) in introductory and intermediate level courses. The policy resulted in a substantial decrease in average grades in the treated departments. As an unexpected side effect, the lowering of average grades had a significant effect on the evaluation of professors in those treated departments. The percentage of students who strongly recommended their professors fell by about 5 percentage points, and there were statistically significant increases in the *neutral* and *do not recommend* categories. "In short, the results strongly indicate that students were less pleased with their instructors, when the grading policy lowered average grades" (Butcher et al., 2014, p. 200).

Correlational studies

Greenwald and Gillmore (1997a) demonstrated that the grades students expect in a course bias their evaluation of course and instructor. These researchers added to the regular rating form at the University of Washington three items that requested judgments that were at best weakly related to quality of instruction. Students were asked to judge the legibility of the instructor's handwriting, the audibility of the instructor's voice, and the quality of the classroom facilities. Their responses to these items showed clear positive relationships with expected grades within a given class. In contrast—consistent with the assumption that these items are peripheral to the quality of teaching—there was no evidence of a grade-rating correlation in the between-courses analyses. The authors concluded that because "all students in the same classroom saw the same instructor's handwriting, heard the same instructor's voice, and had the same classroom teaching aid, the observation of these within-sections relationship . . . suggest the potency of grade influences

on students ratings" (Greenwald & Gillmore, 1997a, p. 1212).

Further evidence of bias in teaching evaluation comes from the analysis of data from the website, "RateMyProfessors.com" (RMP; Felton et al., 2008). RMP is a popular website—founded in 1999—on which students can anonymously rate their professors. According to the 2015 RMP homepage, students have contributed more than 15 million ratings of 1.4 million professors, associated with over 7,000 colleges and universities across the United States, Canada, and the United Kingdom. Because most universities do not give students access to institutional SET information, they find RMP useful for their course shopping (e.g., Hossain, 2010; Vialta-Cerda, McKeny, Gatlin, & Sandi-Urena, 2015). More than 4 million college students access RMP each month (About RateMyProfessors.com, 2015).

Whereas standard SETs only ask students to evaluate a course and to rate their instructor's quality as a teacher, RMP also requires students to rate their professors on a number of dimensions suspected to bias the teaching ratings. Students are instructed on the website to rate their professors on the following four dimensions: (a) *Easiness*: When rating a teacher's easiness, ask yourself: "How easy are the classes this professor teaches? Is it possible to get an A without much work?" (b) *Helpfulness*: Is the teacher "helpful and approachable"? (c) *Clarity*: "How well does the teacher convey the class topics? Is he clear in his presentation? Is he organized and does he use class time effectively?" (d) *Overall Quality*: "The average of a teacher's helpfulness and clarity ratings" (e) *Hotness*: The sum of the "hot" and "not hot" votes, where hot is valued at +1 and not hot at −1.

Felton et al. (2008) were able to use data for all professors in the United States and Canada with at least 20 student ratings at that time (6,851 professors from 369 institutions). They correlated the ratings on all of these dimensions. For our purpose, the most interesting correlation is the correlation between ratings of quality and easiness. At $r = .62$, the two variables are highly correlated. Students rate courses that allow them to get good grades without having to work hard more positively than courses, where good grades require a great deal of work.

Another interesting finding is that quality is highly correlated with the perceived hotness of an instructor ($r = .64$). To demonstrate the impact of this rating of instructor attractiveness, Felton et al. compared the quality scores of the top 100 most and the bottom hundred least attractive professors. The average quality score of the least attractive professors was 2.14 compared to 4.43 for the most attractive professors.

The high correlation between quality and easiness ratings in the website provides support for the assumption that teaching ratings are biased by grading leniency. Although one could still argue that easiness ratings could be a function of quality rather than the other way round, the way easiness is defined on the website (the ability to get a high grade without having to work hard) makes this interpretation rather unlikely. Furthermore, student raters could make comments about the professor to justify their ratings and "professors with high 'Easiness' scores usually received student comments regarding a light workload and high grades" (Felton et al., 2008, p. 40). These comments also rule out the possibility that the hotness ratings were a function of quality rather than the other way round. Student comments about hotness tended to focus on physical characteristics of their professors making it rather unlikely that professors were considered sexy because of their academic brilliance rather than their looks (Felton et al., 2008).

Although in the world of commerce web-based consumer opinion platforms for the exchange of electronic word-of-mouth (eWOM) have become an accepted source of product information for shoppers and also provide them with the opportunity to offer their own consumption-related advice (Fennis & Stroebe, 2016), such platforms are less accepted among educators. As Davidson and Price (2009) criticized, "In a consumerist environment, student evaluations are not 'good' data. They measure how easy the instructor is, how fun, and sometimes, as in the case of the Rate My Professor website, how sexy he or she is. Such data should not be used by students or organizations to evaluate an instructor's ability to teach" (p. 62). A more methodological criticism is that, as a volunteer site, RMP is likely to attract students from the ends of the rating distribution. The information provided on this website could therefore be biased (or more biased than SET ratings).

Surprisingly, studies that correlated institutional SET ratings of faculty members with ratings they received on RMP indicate a great deal of correspondence (M. J. Brown, Ballie, & Fraser, 2009; Sonntag, Bassett, & Snyder, 2009; Timmerman, 2008). M. J. Brown et al. (2009) randomly selected 312 Brooklyn College instructors from the fall 2005 teacher evaluations for the comparison. They selected three items from the 23 questions asked on the Brooklyn College SET that most closely represented the three variables used on RMP (clarity, helpfulness, and easiness) and found moderately high correlations for helpfulness and clarity ($r = .50$ and $r = .59$, respectively) but a lower (though still significant) correlation for easiness ($r = .32$.). The three equivalent SET items were "Teacher's availability to students outside class," "Teacher's ability to communicate clearly," and "Rate the difficulty of examinations in this class" (M. J. Brown et al., 2009, p. 91).

A similar degree of correspondence was reported by Sonntag et al. (2009), who compared SET ratings of 126 professors at Lander University with RMP ratings.

Sonntag et al. found overall quality on RMP to correlate with the equivalent SET ratings ($r = .69$). Most interesting, however, easiness ratings on RMP correlated with the actual GPA of the sample of instructors ($r = .44$). As Sonntag et al. pointed out, these correlations are in the same range as those reported in validity studies of SET that compare SET ratings with evaluations of teaching by administrators (e.g., Kulik & McKeachie, 1975; McKeachie, 1979). Finally, Timmerman (2008), who collected data from 1,167 faculty members from five different universities, reported similarly high correlations.

An apparently discrepant finding was reported by Legg and Wilson (2012), who based their conclusion (stated in their title) that "RateMyProfessors.com offers biased evaluation" (p. 89) on the finding that instructors received significantly lower clarity ratings on RMP than on SET (3.46 vs. 4.04). However, there was no difference on helpfulness ratings, and ratings of easiness were more positive on RMP. Because SET ratings were based on a specific course of an instructor, whereas RMP ratings reflected an overall evaluation, such mean differences are not surprising. More informative would have been the correlations between the two sets of ratings, which were not reported in the article.

It is interesting to note that outside of the world of education, RMP enjoys a much more positive image. Time Magazine recognized the website as one of the best sites of 2008 (50 Best Websites 2008, 2008). Forbes uses RMP ratings as a measure of student satisfaction in their ranking of America's best universities. As the Forbes article "Ranking America's Top Colleges 2015" (2015) states, "Asking students what they think about their course is akin to what some agencies like Consumers Report or J.D. Powers and Associates do when they provide information on various goods or services."

Testing different theories of bias

The few studies that assessed the validity of different theories of bias in teaching evaluation focused mainly on the validity of the attributional account (Gigliotti & Buchtel, 1990; Greenwald & Gillmore, 1997a; Theall, Franklin, & Ludlow, 1990). Their findings are not very conclusive. Theall et al. (1990) reported that external attributions were fewer than expected when the expected grade was an A but more than expected when the expected grades were C, D, or F. This pattern is supportive of the attributional account. In contrast, Gigliotti and Buchtel (1990) reported mixed results. They found that overall students failed to show the pattern of self-serving biases predicted by the self-esteem model. Finally, Greenwald and Gillmore (1997a) rejected the attributional account because it did not predict the negative association between expected grades and workload they found in their studies. Because the attributional theory is concerned with the interpretation of actual rather than expected grades, this finding is not really relevant to that theory. Given the established relationship between physical attractiveness and liking (e.g., Stroebe, Insko, Thompson, & Layton, 1971), the strong association between hotness of an instructor and rating of teaching quality in RMP is consistent with balance theory.

Conclusions

Even though the findings of experimental studies are less than conclusive, they provide tentative evidence that expected or actual grade information biases student evaluation of teaching. Further evidence of bias comes from the findings of Greenwald and Gillmore (1997a) that expected grades biased judgments that were at best weakly related to quality of instruction. However, the strongest evidence of bias is provided by the Felton et al. (2008) finding that perceived grading leniency (i.e., easiness) influences teaching ratings. This finding is consistent with results (reported later) that perceived grading leniency is positively correlated with teaching ratings (Griffin, 2004; Olivares, 2001). Thus, there is strong evidence that the positive association of grades and teaching rating is (at least partly) due to bias.

Unfortunately, research on the theoretical interpretation of this bias is less conclusive. Given the ample support, however, for Weiner's (1979) attributional account of how people interpret success and failure (Zuckerman, 1979) and given that there was some supportive evidence in studies of teaching ratings (Theall et al., 1990), one cannot reject the attributional predictions of the impact of receiving poor grades on students' evaluation of teaching. However, attribution theory would not predict the strong effects of teachers' attractiveness on evaluations of teaching quality found in RMP ratings (Felton et al., 2008). This association is most consistent with an interpretation in terms of balance theory. I offered an integration of balance and attribution theory that would explain these patterns.

4. Assessing Teaching Effectiveness

The evidence reviewed so far provides support for a bias interpretation of the association of grades and teaching evaluation. However, as argued earlier, that there is bias does not rule out the possibility that teaching ratings also reflect the quality of teaching. Because the research reviewed earlier indicates that course grades are a contaminated measure of student learning, other indicators of learning are needed as measure of teacher effectiveness. Some researchers have argued that students' self-reported acquisition of competence could be used as a criterion

(e.g., Machina, 1987). However, students' perception of learning might not always reflect actual learning. Students might think they had learned a great deal in a course when actually they had not (Spooren et al., 2013). This is particularly likely for less competent individuals, who have been shown to overestimate their ability and performance (Kruger & Dunning, 1999). Alternatively, they might use course grades as an indicator of learning. Standardized examinations are also not very practical, because of the large number of courses taught and the emphasis different instructors might make to the same course (Johnson, 2003).

As "a practical solution to this conundrum," Johnson (2003, p. 153) suggested use of student performance in follow-on courses as measure of student learning. He argued that "the simplest measure of teaching effectiveness in a first semester calculus course is the preparation of students for second semester calculus, and average student performance in intermediate Spanish is an obvious measure of the effectiveness in introductory Spanish" (Johnson, 2003, p. 154). The obvious precondition for this procedure is that success in subsequent course requires knowledge and skills gained in the preceding course. By now five studies have related teaching ratings in a concurrent course to student performance in subsequent courses (Braga et al., 2014; Carrell & West, 2010; Johnson, 2003; Weinberg et al., 2009; Yunker & Yunker, 2003). None of those studies has focused on psychology.

Yunker and Yunker (2003) related student achievements, as measured by their grades in intermediate accounting, to their teaching evaluation of introductory accounting (the prerequisite course). The study was based on a sample of 283 students. For each student, the authors had access to grades in the two accounting courses. However, because teacher evaluations are typically given anonymously, the rating used as the predictor variable was the mean rating of a teacher applied by the entire introductory accounting class in which the student had been enrolled. In line with most previous research (Wright & Jenkins-Guarnieri, 2012), course mean ratings of the section of introductory course taken by a student were significantly positively related to the students' grades in that course. However, whereas the association of course ratings and grades in the subsequent course was practically zero, it became significantly negative after controlling for student ability with three variables (GPA, students' grade in the introductory course, ACT score). Because individual grades in the introductory course are highly correlated with both the evaluation of the introductory course and the grades in the subsequent course, controlling for course grade (and other indicators of ability) was essential. However, it is worth pointing out that even the absence of a correlation between evaluation of the introductory course and performance in the subsequent course would be inconsistent with the teacher effectiveness theory.

A second study was conducted at Ohio State University with students who took principles of microeconomics, principles of macroeconomics, and intermediate microeconomics between 1995 and 2004. The data cover more than 45,000 enrollments in almost 400 offerings of these courses. The evaluation instrument contained 10 items, including an overall score, which was the principal measure of student evaluation used in the study. Other questions included measures of perceived learning, the instructor's preparation and organization, and the extent to which students found the course stimulating. As in all previous research, course ratings were positively associated with the grades in the concurrent course. However, when course evaluation was used as a predictor of student performance in subsequent courses (controlling for current grades) no association was found. That this was also true for the measure of student learning (i.e., "learned greatly from the instructor") suggests "that students are not able to evaluate the amount they learn in a course or that they base their beliefs on the grades they expect to receive" (Weinberg et al., 2009, p. 240).

Carrell and West (2010) conducted a study with students at the U.S. Air Force Academy. The great advantage of this study over previous research is that students were randomly assigned to professors and courses, so that results were not affected by selection effects. The data set consisted of 10,534 students who attended the academy from fall 2000 through spring 2007. As in previous research, students' evaluation of professors was positively correlated with grades in the concurrent course. However, when grades in follow-on courses were used as criterion of learning, student evaluations of a concurrent course were significantly negatively correlated with those grades. Carrell and West (2010) concluded that their "results show that student evaluation reward professors who increase achievement in the contemporaneous course being taught, not those who increase deep learning" (p. 430). These authors also reported that students of a less experienced and less qualified professor received significantly better grades on the contemporaneous course but did more poorly in follow-up courses. In contrast, students of more experienced professors showed the reverse pattern.

That these processes are not limited to the United States has recently been demonstrated in a study at Bocconi University, an Italian private university that offers degree programs in economics, management, public policy, and law (Braga et al., 2014). The study is based on the 1998–1999 freshmen, who were randomly allocated to classes. Again, grades in subsequent courses were used as an indication of learning. Student ratings covered

various aspects of teaching, such as lecturing clarity, interest generated by teacher, course logistics, course workload, and an overall rating of teaching quality. The results of this study are consistent with the pattern of findings reported by Yunker and Yunker (2003) and Carrell and West (2010). As in these earlier studies, student ratings of teaching were positively associated with grades in concurrent courses (Braga et al., 2014). In contrast, when performance in future courses was used as criterion of learning, teacher evaluations showed a negative association. As Braga et al. (2014) concluded, "teachers, who are more effective in promoting future performance receive worse evaluation from their students. This relationship was statistically significant for all items of the rating instrument, (except for ratings of course logistics), and was of sizeable magnitude" (p. 81).

The study of Johnson (2003) replicated the negative association between teacher ratings and grades in subsequent courses for some items of the SET but not for others. For example, students' perception of instructors' knowledge, course organization, and course difficulty was negatively related to performance in future courses. On the other hand, self-reported class attendance was positively related to subsequent grades. Another positive predictor was ratings of grading stringency. The more stringent students felt they were graded in a previous course, the better they did in subsequent courses. Consistent with this finding, the average course grade in the prerequisite course, which can be considered an objective indicator of grading stringency, was also negatively related to grades in subsequent courses. This indicates that "courses in which instructors grade more stringently are more effective in preparing students for advanced course" (Johnson, 2003; p. 160). Finally, many of the traditional items of teaching evaluation were unrelated to future grades. Thus, the pattern of results of the study of Johnson (2003) is partly consistent with Weinberg et al. (2009) but also with the findings reported by of Yunker and Yunker (2003) and Braga et al. (2014).

How can we explain the negative association between teaching ratings and student learning, when it is based on their future grades? The most plausible interpretation is grading leniency. Because students seem to be unable to judge what they learned in a course (Weinberg et al., 2009), they might overestimate the amount they learned by basing their estimate on the grade they received. Teachers may also have been rewarded for asking little work from their students and assigning them good grades for inferior performance. Finally, students might have been reinforced for lazy study habits, which were then punished in a subsequent course by a less lenient instructor. As Braga et al. (2014) concluded, "good teachers are those, who require their students to exert effort; students dislike it, especially the least able ones, and their evaluations reflect the utility they enjoyed from the course" (p. 85).

However, it is unclear to what extent the findings reported by Carrell and West (2010) or Braga et al. (2014) can be attributed to grading leniency, because the grading procedures in both schools would seem to prevent such leniency. The same material was taught within a given program, and exam questions were the same for all students. Carrell and West (2010) and Braga et al. (2014) suggested additional mechanisms through which grades could be inflated. Instructors might boost grades by teaching "to the exam." Instead of giving students a deeper understanding of their field and requiring them to do a great deal of reading and writing, they might achieve good exam results by focusing on the specific knowledge that is necessary for answering the exam questions. Poor instructors might also have been less insistent that students do their reading. In social psychology, instructors can tell interesting stories by stringing together the many fascinating (and often counterintuitive) findings and relating them to students' everyday experience. Students are often less interested in learning about the theories behind these studies and the methods used in testing these theories. Yet, learning not only about discoveries but also about the methods of discovery might be a better preparation for future courses. Rather than improving their knowledge of facts, it might improve their ability to think critically.

Regardless of the processes that were responsible, the findings reviewed in this section cannot be reconciled with the teacher effectiveness theory. If the positive association between course grades and concurrent teaching evaluation were a mere reflection of teacher effectiveness, then students in courses that resulted in above-average teaching ratings should do better in future courses. The fact that average teaching ratings in concurrent courses were negatively correlated with grades in subsequent courses is inconsistent with this assumption and suggests that ratings were biased by grades and by other teacher behavior that is liked by students but not conducive to student learning (e.g., use of easy textbooks, showing of many films).

5. Evidence for Grading Leniency

That students' ratings of teaching are biased endangers their validity, but it does not result in grade inflation. In fact, grade inflation would be possible even without student bias, as long as instructors believed that such bias existed. The impact of SETs in inducing grading leniency is determined by instructors' perception of bias and not by actual bias. The fact that grades account for only 10% of the variance in students' teaching evaluations (Wright & Jenkins-Guarnieri, 2012) is therefore no direct indication

for the effect they might have in inducing grading leniency. The main arguments made in this article are that (a) grading leniency is a major cause of grade inflation and (b) one reason for grading leniency is the strategy to "buy" positive teaching evaluations in exchange for assigning good grades without asking for great time investment. This section reviews evidence that most teachers believe that assigning good grades and not requiring too much work has a positive effect on students evaluation of their teaching. In support of the validity of these beliefs, research also shows that grading leniency results in better teaching evaluation. Furthermore, students are more likely to choose courses that are graded leniently than strictly graded courses, which is an additional encouragement for instructors to grade leniently.

Teachers' theories about determinants of teaching evaluation

There are surprisingly few studies of teachers' beliefs about the association of course grades and the demands they impose on students to get good grades on the evaluation of their teaching. Birnbaum (2000) conducted a survey of faculty opinions at the California State University (Fullerton). The survey was sent to all faculty members and was completed by 208 members, who ranged in teaching experience from <12 years to >24 years. Two-thirds of respondents (65.4%) believed that raising standards for grades in their class would result in lower student evaluations of their teaching. Practically the same majority (65.9%) was of the opinion that increasing the amount of content material in their classes would decrease student evaluation of their teaching. Nearly half of the respondents also reported that they now presented less material in their classes than they used to, even though they believed that increasing the content of their course material and of required reading would increase student learning. Birnbaum (2000) drew the following conclusion from his study:

> Our incentive system has produced a decline in standards that diminishes education. Students are motivated to get good grades, and faculty are motivated to get good evaluations. Unfortunately, both these interests can be satisfied by reduction in content and grading standards, which diminishes education. (p. 4)

That those beliefs can influence teaching behavior was demonstrated in a survey conducted at the University of Wisconsin (La Crosse) 4 years after SETs had been introduced (Ryan et al., 1980). The response rate to the questionnaire that was distributed to all 300 faculty members was 63%. Twenty-two percent of respondents indicated that the introduction of SETs has led them to decrease the amount of material covered in their courses and 38% reported lowering the difficulty level of their course. Although the introduction of SETs also resulted in some desirable and appropriated changes in instructional behavior (increased identification of course objectives, 40%; provision of handouts and other course aids, 32%; and use of audio or visual aids, 22.3%), the authors concluded that over all the introduction of SETs "may have had more adverse than positive effects on faculty instructional performance" (p. 329). The authors also found that 93% of their participants believed that faculty morale had decreased somewhat or even greatly as a result of the introduction of SET, with 44% indicating a decrease in their own satisfaction from teaching.

In a web-based study Simpson and Siguaw (2000) sampled members of the Academy of Marketing Science to respond to a questionnaire about SET. Their survey had an extremely low response rate of 9% (52 respondents). Respondents were asked to list activities, which they thought their colleagues had used expressively to influence student evaluation of teaching and evaluate these techniques in terms of their effectiveness. The most frequently mentioned technique was grading leniency (23.6%). "Easy or no exams, unchallenging course material, no required papers, retakes of exam, curving grades, and 'spoon feeding' students with lots of information about the examination" (p. 207) were typical responses in this category. Another frequently mentioned technique was inducements, such as serving cookies, snacks, pizza, and other refreshments on the day evaluations are administered. That this tactic can be effective has been demonstrated in an experiment, in which a person unrelated to the course gave students in half the classes chocolate bars just before they had to respond to a SET questionnaire (Youmans & Jee, 2007). This tactic resulted in a substantial improvement of course ratings in the experimental as compared with the control group.

Despite the fact that information about the beliefs and behaviors of faculty members in response to the use of SET is based on small samples of convenience, the evidence from these studies is quite consistent. The majority of faculty members, who responded to the survey of Birnbaum (2000), believed that raising grades and lowering the workload of students were effective means of improving students' evaluation of their teaching. In addition, a quarter of respondents to the survey of Ryan et al. (1980) acted on these beliefs. The next sections will provide evidence that shows that these beliefs are not unfounded.

The impact of perceived grading leniency on teaching evaluation

The importance of grading leniency as a determinant of teaching evaluation was first pointed out by Greenwald and Gillmore (1997a, 1997b), who collected SET survey

data in more than 500 undergraduate courses at the University of Washington between 1993 and 1994. In addition to the usual items asking evaluations of course and instructor, Greenwald and Gillmore added questions about the expected grade for a course and about workload (i.e., number of hours students spend on that course). One would predict that students would work harder in courses for which they expected to get high rather than low grades. Unexpectedly, the authors found in several studies that students reported doing less work in courses where they expected high grades. Even more interesting, this relationship became stronger when a measure of relative expected grade was used. With this measure, students rated the grade expected in a course against their average grade in other courses. Greenwald and Gillmore (1997a) concluded that only "the leniency theory readily explains the observed negative relationship. The explanation is that strict-grading instructors induce students to work hard in order to avoid very low grades" (p. 1214).

Their findings were replicated at the University of California, San Diego with nearly 8,000 classes covering the years 2003–2007 (Babcock, 2010). Again, the grades students expected in a class were related to the hours they spent outside of class studying for this course. In fact, higher expected grades elicited lower study time, whether the grade expectations were held regarding courses, instructors, or even departments. Babcock (2010) estimated "that a one-point increase in expected grade may reduce weekly study time by about 0.94 hours" (p. 992). A comparison of within- and between-class analyses is particularly interesting. Although there was also a negative correlation between study time and work investment within class, the effect was much smaller than the comparison between classes, suggesting that the association is not due to individual optimism but to the perception of differences in course-specific practices (e.g., grading leniency).

Two further studies, which measured perceived grading leniency directly, were consistent with these findings (Griffin, 2004; Olivares, 2001). Olivares (2001) conducted a survey of 149 students in seven sections of two undergraduate courses (statistics and organizational psychology). Grading leniency was measured with the following question: "Compared to all other college instructors you have had, how would you rate this instructor's grading?" The response scale varied from 1 (*much easier/lenient grader*) to 7 (*much harder/strict grader*). Perceived grading leniency correlated with a global rating of the instructor (r = .42) and with the multiple-items SET scale (r = .45), accounting for 20% of the variance in the ratings of the multiple-items SET used by Olivares. A somewhat lower correlation (r = .21) between perceived grading leniency and teaching evaluation was reported by Griffin (2004), who replicated the Olivares study with a sample of 754 undergraduate students enrolled in 39 education courses. Thus, the findings of Olivares (2001) and Griffin (2004) are consistent with the conclusion that grading leniency is an important determinant of teaching evaluation.

The impact of perceived grading leniency on course selection

If students prefer courses in which they can expect good grades, they should actively select such courses to improve their GPA. This type of biased selection has been demonstrated in several studies (Bar, Kadiyali, & Zussman, 2009; Johnson, 2003; Sabot & Wakeman-Linn, 1991). The earliest study was conducted by Sabot and Wakeman-Linn (1991) at Williams College, a liberal arts college in Massachusetts. These authors studied (in a sample of 376 students) the probability that a student took a second course in a department as a function of the grade he or she received in the first course. Of the male students in economics who did not intend to major in economics (the large majority), the probability of taking a second course in economics was 18.2% less if they received a B rather than an A and 27.6% less if they received a C in the introductory economics course. Of those who did not intend to major in English and were male (again the large majority), the probability of taking a second course in English was 14% less if they received a B rather than an A and 20.3% less if they received a C. Although these findings are consistent with the hypothesis that students prefer courses in which they can expect good grades, there are obviously other factors that could have influenced their decision (e.g., perceived lack of competence).

A longitudinal study with a large sample of students at Duke University provided less ambiguous evidence of the influence of grading leniency on course choice. On the basis of that survey, Johnson (2003) had information about the extent to which students informed themselves about the mean course grades of courses taught in past semesters. Johnson then assessed the influence that grade information, which students had accessed, had on the courses they subsequently took. He found a substantial influence of grade information on choice of future courses. For example, if a student had a choice between courses taught by two instructors, one course having a mean grade of A–, the other having a mean grade of B, the odds that the student would choose the first over the second course was 2 to 1.

In another study, Bar et al. (2009) made use of a policy change at Cornell University, where from 1998 onward, median course grades for all courses were published on a website. Bar et al. used this change to test two

hypotheses: first, that the availability of online grade information would lead to increased enrollment into leniently graded courses and second, that the preference for leniently graded courses would be moderated by student ability with high-ability students being less attracted to the leniently graded courses than their peers. To assess the influence of the newly provided information about average course grades, Bar et al. compared the percentage of students enrolled in courses with a median grade of A and courses with a B median before the new policy (1990–1997) and after the new policy (1998–2004). They found that enrollment in courses with an A median increased from 23.5% to 33.4%, whereas enrollment in courses with a B median decreased from 75.7% to 66.3%. The study also provided evidence for student awareness of the course information. The authors reported that the daily number of visits to the grade information website nearly doubled during periods when students had to enroll for courses compared with other periods of the academic year.

It is important to note that the finding that the preference for leniently graded courses appears to be mainly characteristic for less able students; that is, students with lower SAT scores. Therefore, even though Johnson's (2003) conclusion that "an instructor who grades stringently is not only less likely to receive favorable course evaluations, but is also less likely to attract students" (Johnson, 2003, p. 193) is correct in general, a tough grading instructor who is also a good teacher might be rewarded by attracting mainly good students who are interested in learning about their discipline rather than in merely getting good grades.

The dark side of grading leniency

One could argue that grading leniency is a win-win situation: Students receive better grades than they deserve, which advances their job prospects, and teachers receive better student evaluations than they deserve, which improves their standing within the faculty and increases their chances to get tenure. Unfortunately, there is a dark side to grading leniency. It is likely to demotivate students. The findings of Babcock (2010) and Greenwald and Gillmore (1997a, 1997b) that students reported doing more work in courses that had low expected grades than that that had high expected grades supports this assumption. Furthermore, as Rosovsky and Hartley (2002) argued, grades are intended to inform students of their strengths, weaknesses, and areas of talent, which will help them in their career choice. With grade inflation, grades have become less indicative for the students themselves.[1] Sabot and Wakeman-Linn (1991) found that the grades students received in a department that graded strictly were much more indicative of future performance than grades received in a department that graded leniently. They reported that in a department that graded strictly the correlation between the grades a student received in a first course with that in a second course was .61, compared with .37 in a leniently grading department. Furthermore, studies that related the average course grade to students' performance in subsequent courses found a negative association (Johnson, 2003): Students who took courses with an instructor who graded leniently did less well in subsequent courses. Given the findings about the negative association of expected course grade and time investment, these findings are not unexpected (Babcock, 2010; Greenwald & Gillmore, 1997b).[2] Lenient grading and easy courses are also likely to discourage bright and achievement-motivated students. If students end up with the same grade as their less able and less motivated fellow students, they might decide that investing time in their study is not worth the effort. Finally, lenient grading invalidates grades as selection criteria on the job markets. Firms trying to select the most able students will have to look for other sources of information in selecting employees.

A grading system where the median grade is A– can have damaging effects and should be changed. As the grading policy adopted by Wellesley College in 2000 shows, such changes can be effective. The problem is that unless such changes are adopted by all schools within the same segment (e.g., Ivy League), such changes put the college that introduces them at a competitive disadvantage. This was probably the reason for the grade deflation reversal at Princeton, a university that in 2004 had instituted a policy that prescribed for every department and program that A-range grades (A+, A, A–) were to account for less than 35% of the grades given in undergraduate courses. Although the policy was effective in rebalancing the grading scale, Princeton rescinded it in 2014. According to a report in the *Daily Princetonian*, "At the time of the policy's implementation, no peer institutions followed the University's lead in taking institution-wide measures to curb grade inflation, prompting criticism that the policy could hurt students when applying for positions post-graduation" (Windemuth, 2014). Harvard economics professor J. A. Miron agreed with this view, stating that "the policy put Princeton in a tough position because some students concerned about their grades would tend to choose other schools over Princeton". . . "Unless other schools followed suit, it was a competitive mistake" (Bernhard, 2014).

A final question

A final question that needs to be addressed is how to explain the differences in grade inflation between public and private universities and between science departments

and arts, humanity, and social science departments at all universities (Rojstaczer, 2015). The greater grade inflation at private schools could be due to the need of these schools to work harder to achieve consumer satisfaction, given that the fees they charge are considerably higher than those of public universities. There is likely to be greater pressure on instructors to achieve good teaching ratings both from the university administration and from students, who may consider themselves elite students, who deserve top grades.

That science departments grade on average roughly 0.4 points lower on a 4.0 scale than humanities departments and 0.2 points lower than social science departments (Rojstaczer & Healy, 2010) could be due to differences both in the expectations students hold with regard to the function of lectures and course work and to differences in the beliefs faculty holds about the function of grades. It would seem plausible that students of physics or mathematics are less likely than students in humanities and social science departments to expect lectures to be entertaining, be relevant to everyday life, or advance their personal growth. Physics and mathematics students probably expect lectures to help them to understand the difficult material they are studying. These expectations match the teaching goals of the respective disciplines. There is evidence that "science and mathematics faculty are more concerned with teaching facts and principles of their disciplines, whereas faculty in the arts were more likely to view their primary teaching role as fostering student development and personal growth" (Barnes, Bull, Campbell, & Perry, 2001, p. 456/7).

With regard to faculty beliefs about the function of grades, Goldman (1985) deplored in an article titled "The betrayal of the gatekeepers: Grade inflation," that most university educators have abandoned their role as the gate keepers who help society to sort people according to their abilities. Gatekeepers are instructors who believe that "a high percentage of A's in a class indicates low standards or a lack of rigor in assessing achievement" (Barnes et al., 2001). Instructors, who see their role as gatekeepers are likely to be resistant to the pressures toward grading leniency. Barnes et al. (2001) suggested that science teachers are more likely than instructors from humanities or social sciences to believe in the gatekeeping function of grades. In a survey of a sample of 442 faculty members from different disciplines who responded to a newly developed "Beliefs about Grades Inventory," Barnes et al. (2001) found indeed that faculty "in hard, pure, nonlife disciplines (e.g., physics) had significantly higher mean scores on a measure of 'gatekeeping' than faculty in soft pure nonlife disciplines (e.g., history)" (p. 484). Although more evidence is needed, it seems plausible that differences in student attitudes toward the function of lectures and in faculty beliefs about the function of grade could be responsible for the apparent ability of science department to avoid grade inflation.

Conclusions

This article addressed the paradox that there is an increase in college and university students' grade point average since the 1980s, even though students appear to spend less and less time on their studies. I presented evidence that the university policy of using teaching evaluations for important decisions such as salary increases, tenure decisions, promotions, or the appointment of new faculty is an important cause of this grade inflation. This policy, which has been nearly universally adopted by U.S. colleges and universities, puts strong pressure on teachers to improve their teaching evaluations. Teachers can achieve this to some extent by trying to improve their teaching and make their lectures more interesting. However, students' ratings of a course and of the instructor are at least partly influenced by the amount of work they are required to do and by the grades they expect or receive.

Because most teachers are aware of this bias, they are confronted with the conflict that if they require students to work hard and award top grades only to students who deserve them, they might not receive very favorable teaching ratings. Furthermore, unless their course is part of the degree requirement (and they teach the only section), they are likely to attract few students. Thus instructors who are competent but not brilliant teachers are faced with the alternative of either foregoing top teaching ratings or lowering the standard in their courses. Whereas established professors who also have a good record as researchers do not have to yield to such pressure, the price of resistance might be too high for young academics who are not yet tenured. This might explain why in some of the studies reported earlier, established professors received less good ratings in concurrent courses but performed well when grades in subsequent course was used a criterion of learning (e.g., Carrell & West, 2010).

The findings reported in this article do not support teacher effectiveness theory and are more consistent with an alternative interpretation proposed by Braga et al. (2014) that student ratings of teaching reflect consumer satisfaction or in economic terms, "realized utility" (p. 84). However, the presence of biases does not preclude the possibility that the quality of teaching also influences consumer satisfaction. After all, a well-structured and informative lecture is likely to be more satisfying than a chaotic presentation that contains little information. Furthermore, the evidence from studies of the convergent validity of student ratings suggests that they provide some indication of instructional quality (e.g., Kulik, 2001;

Marsh, 1984). Even Johnson (2003) observed in his otherwise critical account of student ratings of teaching that in his study "for every item, the best predictor of an individual student's response was the consensus rating of the item by other students" (p. 95). Despite all their weaknesses, student ratings provide some information about what is happening in courses and whether teachers are doing a reasonably good job.[3]

However, one could make student ratings more informative by removing some of the most obvious biases. One way to remove bias would be a statistical adjustment suggested by Greenwald and Gillmore (1997b). Greenwald and Gillmore argued that using the expected grade measures as basis for a covariance adjustment would remove much of the influence of grading policy on student ratings. Administrators could also use additional information for their evaluation of faculty members such as teaching portfolios. In addition, classroom visits by experienced colleagues might not only be informative with regard to teaching quality but also provide helpful advice for instructors to improve their teaching. Finally, one could use the Wellesley approach to prevent grade inflation by instituting a rule about average grades in undergraduate introductory and intermediate courses (Butcher et al., 2014). The problem with that approach is that, unless it is generally adopted, it puts students from Wellesley at a disadvantage.

It would also ease the pressure—experienced particularly by younger faculty—to improve teaching ratings by grading more leniently if administrators deemphasized the importance of SET information. Administrators could explain that teaching evaluations are mainly used to distinguish acceptable from unacceptable teaching. Furthermore, administrators could abandon the cult of celebrating top teachers. There are many ways by which top ratings can be achieved and not all of them will also result in top learning. It is interesting to note that a recent study found a nonlinear relationship between global instructor ratings and their outcome measure of student learning (Galbraith, Merrill, & Kline, 2012). The study was based on 116 business-related courses in a "School of Business" at a private university in the United States. The school had "invested substantial time and resources in revising and quantifying its learning outcome assessment process" for all the core courses (Galbraith et al., 2012, p. 358). As to be expected, students learned least with teachers at the bottom end of the distribution of teaching ratings. More surprisingly, however, top-rated instructors were equally ineffective in achieving student learning. Although this finding needs to be replicated, it should provide comfort to the large majority of university teachers who have failed in their effort to move their teaching evaluation to the top region of the scale.

Declaration of Conflicting Interests

The author declared no conflicts of interest with respect to the authorship or the publication of this article.

Notes

1. Harvey Mansfield, Professor of Government at Harvard, has developed an unconventional solution to this problem. He gives students two grades. The first is the grade that students actually deserve, namely a C for mediocre work, a B for good work, and an A for excellence. This grade will be issued only to the student for every paper and exam. The second grade, computed only at semester's end, will be what Mansfield called an "ironic grade" (*ironic* in this case being a word used to mean lying), and it will be computed on a scale that takes as its mean the average Harvard grade. This higher grade will be sent to the registrar's office and will appear on public transcripts. It will be the public grade and ensure that students are not penalized for taking one of Mansfield's classes (Mansfield, 2012).
2. Consistent with this, a study of students' gains in critical thinking, analytical reasoning, and written communication (measured with the Collegiate Learning Assessment; CLA) over their 4 years in college revealed that 36% of students did not show any significant improvement in learning as measured with the CLA. However, students, who "took courses requiring both significant reading (more than 40 pages per week) and writing (more than 20 pages over the course of the semester) had higher rates of learning; students reporting faculty with high expectations at their institutions had higher rates of learning" (Arum & Roksa, 2011, p. 205).
3. Suggestive evidence for the validity of student ratings of teaching under conditions that remove incentives for bias comes from a study by Wang, Pascarella, Nelson Laird, and Ribera (2015). These researchers asked a sample of students at the end of their 4-year study to evaluate "the extent to which they were exposed to clear and organized instructions across all their classes" (p. 1793). These perceptions correlated moderately positively with improvements in critical thinking skills and need for cognition. Because these ratings are no longer about the teaching of a specific instructor but about teaching quality in general, students would not have been motivated by a specific course grade to give biased ratings. However, if we assume that students, who improved their critical thinking skills and their need for cognition during the 4 years of study might also have received better grades, their evaluations could still have been biased by the glow of success.

References

50 Best Websites 2008. (2008). *Time Magazine*. Retrieved from http://content.time.com/time/specials/2007/article/0,28804,1809858_1809956_1811548,00.html#ixzz1c19iML00

About RateMyProfessors.com. (2015). Retrieved from http://www.ratemyprofessors.com/About.jsp

Arum, R., & Roksa, J. (2011). Limited learning on college campuses. *Society*, *48*, 203–207.

Babcock, P. (2010). Real costs of nominal grade inflation? New evidence from student course evaluations. *Economic Inquiry*, *48*, 983–996.

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*, 468–478.

Bar, T., Kadiyali, V., & Zussman, A. (2009). Grade information and grade inflation: The Cornell experiment. *Journal of Economic Perspectives*, *23*, 93–108.

Barnes, L. L. B., Bull, K. S., Campbell, J., & Perry, K. M. (2001). Effects of academic discipline and teaching goals in predicting grading beliefs among undergraduate teaching faculty. *Research in Higher Education*, *42*, 455–467.

Bernhard, M. P. (2014, October 9). Princeton grade deflation reversal disappoints some here. *The Harvard Crimson*. Retrieved from 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*. Retrieved from http://psych.fullerton.edu/mbirnbaum/faculty3.htm

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

Brown, D. L. (1976). Faculty ratings and student grades: A university-wide multiple regression analysis. *Journal of Educational Psychology*, *68*, 573–578.

Brown, M. J., Ballie, M., & Fraser, S. (2009). Rating RateMyProfessors.com: A comparison of online and official student evaluations of teaching. *College Teaching*, *57*, 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*, 189–204.

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

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, *44*, 495–518.

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

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

Cohen, P. A. (1982). Validity of student ratings in psychology course: A research synthesis. *Teaching of Psychology*, *9*, 78–82.

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

Davidson, E., & Price, J. (2009). How do we rate? An evaluation of online student evaluations. *Assessment & Evaluation in Higher Education*, *34*, 51–65.

Dowell, D. A., & Neal, J. A. (1982). A selective review of the validity of student ratings of teaching. *Journal of Higher Education*, *27*, 459–463.

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*, 45–61.

Fennis, B. M., & Stroebe, W. (2016). *The psychology of advertising* (2nd ed.). London, England: Routledge.

Galbraith, C. S., Merrill, G. G., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education*, *53*, 353–374.

Gigliotti, R. J., & Buchtel, F. S. (1990). Attributional bias in course evaluations. *Journal of Educational Psychology*, *82*, 341–351.

Goldman, L. (1985). The betrayal of the gatekeepers: grade inflation. *The Journal of General Education, 37*, 97–121.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, *25*, 161–178.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings. *American Psychologist*, *52*, 1182–1186.

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

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*, 743–751.

Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instructions. *Contemporary Educational Psychology*, *29*, 410–425.

Guthrie, E. R. (1953). The evaluation of teaching. *American Journal of Nursing*, *53*, 220–221.

Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: Wiley.

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology*, *68*, 130–133.

Hossain, T. M. (2010). Hot or not: An analysis of online professor-shopping behavior of business students. *Journal of Education for Business*, *85*, 165–171.

Jacobsen, E. (2015). *Average SAT scores of college-bound seniors (1952–present)*. Retrieved from http://www.erikthered.com/tutor/historical-average-SAT-scores.pdf

Jewell, R. T., McPherson, M. A., & Tieslau, M. A. (2013). Whose fault it? Assigning blame for grade inflation in higher education. *Applied Economics*, *45*, 1185–1200.

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

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

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-awareness. *Journal of Personality and Social Psychology*, *77*, 1121–1134.

Kulik, J. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research*, *109*, 9–25.

Legg, A. M., & Wilson, J. H. (2012). RateMyProfessors.com offers biased evaluations. *Assessment & Evaluation in Higher Education*, *37*, 89–97.

Machina, K. (1987). Evaluating student evaluations. *Academe*, *73*, 19–22.

Mansfield, H. (2012). *Harvard Harvey Mansfield 3 of 5- grade inflation average Harvard*. Available from https://www.Youtube.com

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility. *Journal of Educational Psychology*, *76*, 707–754.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for further research. *International Journal of Educational Research*, *11*, 253–288.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workloads on students' evaluations of teaching: Popular myth, bias, validity or innocent bystanders? *Journal of Educational Psychology*, *92*, 202–208.

McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education*, *21*, 150–158.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe*, *65*, 384–397.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist*, *52*, 1218–1225.

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

Olivares, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology*, *26*, 382–399.

Pascarella, E. T., Blaich, C., Martin, G. L., & Hanson, J. M. (2011). How robust are the findings of Academically Adrift? *Change*, *43*, 20–24.

Ranking America's top colleges 2015. (2015, July 29). *Forbes*. Retrieved from http://www.forbes.com/sites/carolinehoward/2015/07/29/ranking-americas-top-colleges-2015/#2715e4857a0b267e6aaa1b04

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

Rojstaczer, S. (2015). *Grade inflation at American colleges and universities*. Available from 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*, ID Number: 15928. Available from http://www.tcrecord.org

Rosovsky, H., & Hartley, M. (2002). *Evaluation and the academy: Are we doing the right thing?* Academy of Arts & Sciences. Retrieved from 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*, 317–333.

Sabot, R., & Wakeman-Linn, J. (1991). Grade inflation and course choice. *Journal of Economic Perspectives*, *5*, 159–170.

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*, 99–213.

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*, 499–504.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83*, 598–642.

Stroebe, W., Insko, C. A., Thompson, V. D., & Layton, B. D. (1971). Effects of physical attractiveness, attitude similarity, and sex on various aspects of interpersonal attraction. *Journal of Personality and Social Psychology*, *18*, 79-91.

Theall, M., Franklin, J., & Ludlow, L. H. (1990, April). *Attributions or retributions: Student ratings and the perceived cause of performance*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA. Retrieved from http://files.eric.ed.gov/fulltext/ED319764.pdf

Timmerman, T. (2008). On the validity of RateMyProfesors.com. *Journal of Education for Business*, *84*, 55–61.

Vasta, R., & Sarmiento, R. F. (1979). Liberal grading improves evaluations but not performance. *Journal of Educational Psychology*, *71*, 207–211.

Vialta-Cerda, A., McKeny, P., Gatlin, T., & Sandi-Urena, S. (2015). Evaluation of instruction: Students' patterns of use and contribution to RateMyProfessors.com. *Assessment & Evaluation in Higher Education*, *40*, 181–198.

Wang, J. S., Pascarella, E. T., Nelson Laird, T. F., & Ribera, A. (2015). How clear and organized classroom instruction and deep approaches to learning affect growth in critical thinking and need for cognition. *Studies in Higher Education*, *40*, 1786–1806.

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

Weiner, B. (1979). A theory of motivation for some classroom experiences. *Journal of Educational Psychology*, *71*, 3–25.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York, NY: Springer.

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

Wilson, K. L., Lizzo, A., & Ramsden, P. (1997). The development, validation and application of the Course Experience Questionnaire. *Studies in Higher Education*, *22*, 33–54.

Windemuth, A. (2014, October 6). After faculty vote, grade deflation policy officially dead. *Daily Princetonian*. Retrieved from http://dailyprincetonian.com/news/2014/10/breaking-after-faculty-vote-grade-deflation-policy-officially-dead/

Worthington, A. G., & Wong, P. T. (1979). Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology*, *71*, 764–775.

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education*, *37*, 683–699.

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

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*, 313–317.

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality*, *47*, 245–287.