

# Chapter 15

## Lessons Learned from Research on Student Evaluation of Teaching in Higher Education



Bob Uttl

**Abstract** In higher education, anonymous student evaluation of teaching (SET) ratings are used to measure faculty's teaching effectiveness and to make high-stakes decisions about hiring, firing, promotion, merit pay, and teaching awards. SET have many desirable properties: SET are quick and cheap to collect, SET means and standard deviations give aura of precision and scientific validity, and SET provide tangible seemingly objective numbers for both high-stake decisions and public accountability purposes. Unfortunately, SET as a measure of teaching effectiveness are fatally flawed. First, experts cannot agree what effective teaching is. They only agree that effective teaching ought to result in learning. Second, SET do not measure faculty's teaching effectiveness as students do not learn more from more highly rated professors. Third, SET depend on many teaching effectiveness irrelevant factors (TEIFs) not attributable to the professor (e.g., students' intelligence, students' prior knowledge, class size, subject). Fourth, SET are influenced by student preference factors (SPFs) whose consideration violates human rights legislation (e.g., ethnicity, accent). Fifth, SET are easily manipulated by chocolates, course easiness, and other incentives. However, student ratings of professors can be used for very limited purposes such as formative feedback and raising alarm about ineffective teaching practices.

**Keywords** Student evaluation of teaching · SET · Validity · Teaching effectiveness

### 1 Introduction

In higher education, anonymous student evaluation of teaching (SET) are used to measure the teaching effectiveness of faculty members and to make high-stakes decisions about them, such as hiring, firing, promotion, tenure, merit pay, and teaching awards (Uttl et al., 2017). If available to students, they are also used by students for course selection in the same manner as the popular website [www.ratemyprofessor.com](http://www.ratemyprofessor.com) (RMP). SET have their allure: (a) SET are quick and cheap to administer; (b)

---

B. Uttl (✉)  
Mount Royal University, Calgary, AB, Canada  
e-mail: [buttl@mtroyal.ca](mailto:buttl@mtroyal.ca)

SET means and standard deviations give an aura of precision and scientific validity; and (c) SET provide tangible seemingly objective numbers for high-stake decisions and public accountability purposes. However, a still little known legal case from Ryerson University in Toronto (*Ryerson University v. Ryerson Faculty Association*, 2018 CanLII 58446, available at [www.canlii.org](http://www.canlii.org)) is a wake-up call about the uninformed use of SET, and reminder that SET are not valid as a measure of faculty's teaching effectiveness. In this chapter, I review the evidence against SET, evidence showing that they do not measure teaching effectiveness, vary predictably across factors completely irrelevant to faculty's teaching effectiveness, and can be raised with something as small as a Hershey kiss. I will also argue that the widespread use of SET may be one of the main contributors to grade inflation, driving up grades over the past 30 years, during a time period when time-spent studying has been steadily decreasing and the proportion of high school students entering colleges and universities increasing.

Typically, within the last few weeks of classes, students are asked to rate professors on various scales. A university evaluation unit then summarizes the ratings for each class and, after the classes are over and the final grades assigned, various statistical summaries including means and standard deviations are then provided to faculty and their administrators. These summaries may include departmental, faculty, or university "norms," such as the means and standard deviations of all course means within the department, faculty, and/or university. These summaries are then used as the key, if not sole, evidence of faculty teaching effectiveness (Uttl et al., 2017).

At the same time, no standards for satisfactory SET ratings are provided to anyone. Evaluators—chairs, deans, tenure and promotion committees, provosts, and presidents—use their own individual standards to arrive at their decisions about faculty teaching effectiveness. It is not uncommon for these evaluators to believe that faculty members falling below the mean are unsatisfactory and in need of improving their teaching. Moreover, these evaluators change periodically and unpredictably, even within the typical six-year time frame between a faculty member's initial hiring and eventual decision about promotion and/or tenure.

There are three types of commonly-used SET tools—those that are developed in-house by an institution, those that are obtained for free, such as the SEEQ (Marsh, 1980, 1991), and those that are developed commercially for purchase, such as the ETS SIR-II sold by the Education Testing Service ([www.ets.org](http://www.ets.org)), the IDEA SRI sold by IDEA Center ([www.ideaedu.org](http://www.ideaedu.org)), and the CIEQ sold by C.O.D.E.S Inc. ([www.cieq.com](http://www.cieq.com)). In all of these systems, faculty's SET ratings are often compared to the departmental, faculty, university, or "national norms" (i.e., the average SET ratings for all institutions that purchased a particular commercial SET system). The commercial systems also give some guidelines on interpretation of SET. For example, the C.O.D.E.S. Inc guidelines specify that faculty scoring below the 70th percentile need at least "some improvement," implying that only the top 30% of faculty with the highest SET ratings are good enough and need "no improvement" (see [www.cieq.com/faq](http://www.cieq.com/faq)). Notably, all of the commercial systems are explicitly intended to be used for both faculty development (formative uses) and for high-stakes personnel

decisions (sumative uses) and their developers believe that they are valid measures of teaching effectiveness.

The focus on norm-referenced interpretation of SET ratings, requiring faculty to place above the 30th, 50th, or even 70th percentile, to avoid criticism of their teaching, will always, by definition, result in large proportions of unsatisfactory and in “need of at least some improvement” faculty members. Assuming few faculty members want to be labeled unsatisfactory or in “need of at least some improvement”, this type of norm-referenced interpretation of SET sets up and fuels a race among faculty members to reach as high of ratings as possible. By definition, depending on the specific percentile cut-offs, 30, 50, or 70% of the faculty will lose this race. The higher the percentile cut off, the more intense and more high-stakes the race becomes.

Regardless of the specific percentile cut-offs for “unsatisfactory” or in “need of some improvement” labels, some proponents of SET ratings also argue that SET identify faculty members who successfully match their academic standards, teaching demands, and workload to students’ abilities. For example, in response to arguments that SET are responsible for grade inflation and work deflation, Abrami and d’Apollonia (1999) argued:

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

As Uttl et al. (2017) observed, in Abrami and d’Apollonia’s (1999) view, SET are an appropriate standards meter allowing professors to determine what students’ perceive to be an appropriate workload, appropriate amount to learn for specific grades, and, in short, a proxy of appropriate academic standards from the students’ perspective. Professors who get high SET ratings are appropriately matching their standards to students’ standards and professors who get low SET ratings are failing to do so.

## **2 SET Are an Invalid Measure of Faculty Teaching Effectiveness**

Are SET a valid measure of faculty teaching effectiveness? Do students learn more from more highly rated professors? If SET are a valid measure of faculty’s teaching effectiveness, SET ought to strongly correlate with student achievement attributable to the professors’ teaching styles, and ought not to be influenced by teaching effectiveness irrelevant factors (TEIFs) such as students’ intelligence, cognitive ability, prior knowledge, motivation, interest, subject field, class size, class meeting time, etc. SET also ought not to be influenced by certain students preference factors (SPFs) such as professors’ hotness/attractiveness, age, gender, accent, nationality, ethnicity, race, disability, etc., whose consideration runs afoul to human rights legislation. Finally, SET ought not to be influenced by ill-advised or detrimental to student learning factors

(DSLFS) such as professors reducing workloads, inflating grades, and distributing chocolates and cookies. Review of the literature, however, now convincingly shows that SETs are not a valid measure of teaching effectiveness, that students do not learn more from more highly rated professors and that SET are substantially influenced by numerous TEIFs, SPFs, and DSLFS.

## ***2.1 There Is No Widely Accepted Definition of Effective Teaching***

The first fundamental problem in assessing the validity of SET as a measure of faculty teaching effectiveness is that professors, administrators, and even experts do not agree on what effective teaching is (Uttl et al., 2017). In turn, experts do not even agree on which teaching methods are effective and which specific teaching behaviors amount to effective teaching. For example, some professors, administrators, and experts believe that teaching methods such as unannounced pop quizzes, questioning students in front of their peers, and encouraging student attendance by leaving out words or phrases from lecture slides are effective teaching methods. In contrast, others believe that these same methods are insensitive, anxiety-producing, and even demeaning, disrespectful, and detrimental to student learning.

In the absence of an agreed upon definition, it is impossible to measure effective teaching. However, the experts do agree that effective teaching ought to result in student learning (Uttl et al., 2017). Accordingly, studies attempting to establish the validity of SET as a measure of effective teaching have focused on determining the correlation between professors' mean class SET ratings and student achievement.

## ***2.2 Students Do Not Learn More from More Highly Rated Professors***

For nearly 40 years, the key evidence cited to support the validity of SET as a measure of faculty teaching effectiveness have been multisection studies that examine the correlations between the mean class SET and the mean class student achievement on common exams. An ideal multisection study has several critical features: (a) it examines the correlation between SET and student achievement in a large course split into numerous smaller sections, with each section taught by a different professor, (b) professors follow the same course outline, use the same assessments, and the same final exam, (c) students are randomly assigned to the sections, and (d) SET are administered prior to the final exam at the same time to all sections. In this design, if students learn more from more highly rated professors, the sections' average SET ratings ought to be highly correlated with sections' average final exam scores. Experts have generally agreed that multisection studies are the strongest evidence for

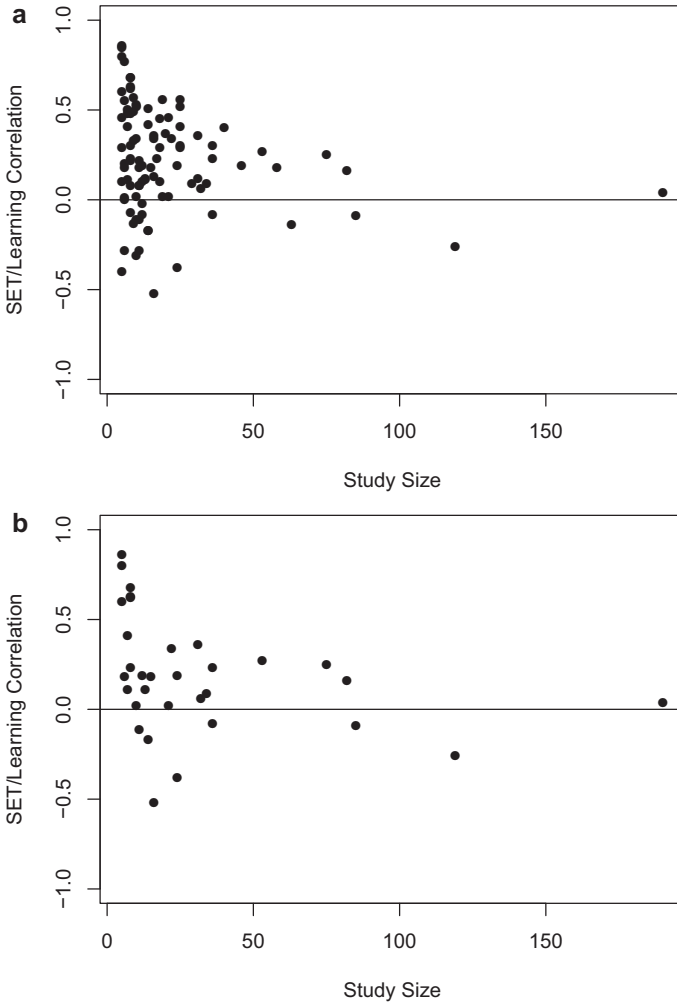
determining the validity of SET as a measure of professors' teaching effectiveness, that is, professors' contribution to students' learning (Uttl et al., 2017).

Cohen (1981) published the first meta-analysis of 67 multisection studies available to that date and reported a small-to-moderate SET/learning correlation  $r = 0.43$ . Cohen concluded: "The results of the meta-analysis provide strong support for the validity of student ratings as a measure of teaching effectiveness" (p. 281) and continued: "we can safely say that student ratings of instructors are a valid index of instructional effectiveness. Students do a pretty good job of distinguishing among teachers on the basis of how much they have learned" (p. 305). Cohen's findings and conclusions have subsequently been cited over 1,000 times as evidence of SET validity as a measure of faculty teaching effectiveness (Web of Science, Google Scholar).

However, Uttl et al. (2017) recently demonstrated that Cohen's (1981) conclusions were unwarranted, and the result of flawed methods and data analyses. Most critically, Cohen disregarded the sample sizes of primary studies in his meta-analysis. In doing so, he gave equal weight to many small sample sized studies as he gave to fewer larger sample sized studies. Compounding this problem, Cohen also failed to take into account small sample size bias clearly visible from scatterplots of SET/learning correlations as a function of sample size. After taking into account small sample size bias, the best estimate of SET/learning correlation was only  $r = 0.27$ , substantially less than  $r = 0.43$  reported by Cohen. Uttl et al. (2017) reported a new updated analysis of 97 multisection studies. Figure 1, Panel A, shows the results of Uttl et al.'s new updated meta-analysis based on 97 multisection studies. It confirms the strong small sample size bias already visible in Cohen's (1981) data set. Taking into account the small sample size bias, the best estimate of SET/learning correlation from this new meta-analysis is  $r = 0.08$ . Panel B shows the Uttl et al. results but only for studies that adjusted the SET/learning correlations for prior learning/ability. The best estimate of SET/learning correlations taking into account both the small sample size bias and prior learning/ability is nearly zero,  $r = -0.02$ . Accordingly, taking into account small sample size bias and prior learning/ability, the multisection studies demonstrate that SET/learning correlations are zero. In other words, students do not learn more from more highly rated professors.

### ***2.3 SET Are Influenced by Many Teaching Effectiveness Irrelevant Factors***

SET correlate with numerous TEIFs such as students' intelligence, cognitive ability, interest, and motivation; subject field; class size; etc.



**Fig. 1** The results of meta-analyses of multisection studies. Panel **A** shows the scatterplot of SET/learning correlations by study size for Uttil et al.'s (2017) new updated meta-analysis. After taking into account a small sample bias, the SET/learning correlation was only  $r = 0.08$  for SET averages. Panel **B** shows Uttil et al. (2017) results but only for studies that adjusted the SET/learning correlations for prior learning/ability. After taking into account both the small sample size bias and prior/learning ability, the SET/learning correlation is nearly zero,  $r = -0.02$

### 2.3.1 Students Intelligence, Ability, and Kruger Dunning Effect

Numerous studies have demonstrated that people are generally very poor in assessing their own cognitive abilities including attention, learning, and memory. Correlations between self-assessment of abilities and performance on objective tests of those

abilities are generally close to zero (Uttl & Kibreab, 2011; Williams et al., 2017). Yet, many SET forms ask students to rate how much they learned from their professors.

Furthermore, as Kruger and Dunning (1999) demonstrated, people's self-assessment of their abilities depends on the abilities themselves. Those scoring low on objective ability tests hugely overestimated their performance whereas those scoring high on objective ability tests tended to underestimate their own performance. Moreover, low-ability individuals were less able to distinguish superior performance from inferior performance of their peers. As Kruger and Dunning observed, the incompetent are not only incompetent but their incompetence deprives them of the ability to recognize their own incompetence as well as the competence of others. It is self-evident that students who believe that their work deserved As or Bs but received Ds or Fs are unlikely to be satisfied and unlikely to give their professors high SET ratings.

### 2.3.2 Student Interest and Motivation

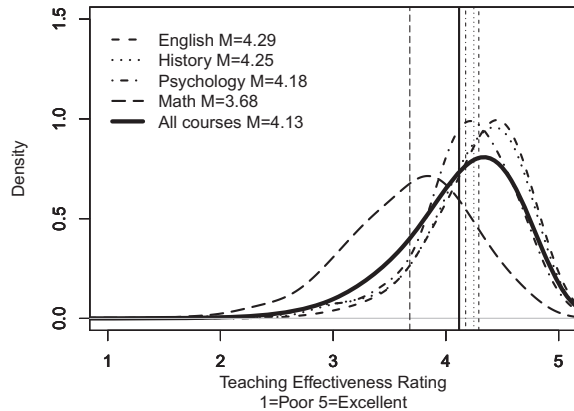
Hoyt and Lee (2002) reported SET ratings by student motivation and class size for the 20 items of the IDEA SRI. Student motivation was measured by a question "I really wanted to take this course regardless of who taught it." Collapsed across questions and class size, the least motivated students gave SET ratings that were 0.44 lower than those of the most motivated students, corresponding to an approximately 0.75 standard deviation difference. Moreover, this effect was substantial on each and every question, ranging from a 0.24 to 0.70 difference on a 1–5 rating scale.

### 2.3.3 Course Subject

Centra (2009) reported that the natural sciences, mathematics, engineering, and computer science courses were rated substantially lower, about 0.30 standard deviation lower, than courses in humanities such as English, history, and languages. Similarly, Beran and Violato (2009) reported that courses in natural science were rated 0.61 standard deviation lower than courses in social science. Surprisingly, Centra as well as Beran and Violato concluded that these effects were ignorable.

Using 14,872 course evaluation data from a US mid-sized university, Uttl and Smibert (2017) demonstrated that the differences in SET ratings between subjects such as English and Math are substantial (the difference between the means was 0.61 on a 5-point scale), and that professors teaching quantitative courses are far more likely to be labeled unsatisfactory when evaluated against common criteria for a satisfactory label. Figure 2 shows the distribution of SET ratings for Math, English, Psychology, History, and all courses. The distributions of math professors ratings are more normal and substantially shifted toward less than excellent ratings whereas the distribution of English, history, psychology, and all professor courses professors ratings are higher and positively skewed. Thus, if the same standards are applied to professors teaching quantitative vs. non-quantitative courses, professors teaching

**Fig. 2** Smoothed density distribution of overall mean ratings for all courses and for courses in selected subjects. Figure highlights that math professors received much lower ratings than professors in English, History, Psychology, and all courses (from Uttl & Smibert, 2017, Fig. 1) (a smoothed density distribution can be thought of as a smoothed histogram with area below the curve equal to 1)



quantitative courses are far more likely to be not hired, fired, not re-appointed, not promoted, not tenured, denied merit pay, and denied teaching awards.

Of course, the simple fact that professors teaching quantitative vs. non-quantitative courses receive lower SET ratings is not evidence that SETs are biased. It may be that professors teaching quantitative vs. non-quantitative courses are simply incompetent, less effective teachers. However, as pointed out by Uttl and Smibert (2017) this incompetence explanation is unlikely. A wealth of evidence strongly suggests that the lower ratings of professors teaching quantitative vs. non-quantitative courses is due to factors unrelated to professors themselves. First, the mathematical knowledge and numeracy abilities of populations worldwide have decreased over the years. For example, half of Canadians now score below the level required to fully participate in today's society (Orpwood & Brown, 2015). Second, Uttl et al. (2013) found that fewer than 10 out of 340 undergraduate students were "very interested" in taking any one of the three statistics courses offered in the psychology department at Mount Royal University. In contrast, 159 out of 340 were "very interested" in taking the Introduction to the Psychology of Abnormal Behavior. Thus, professors teaching statistics classes vs. abnormal psychology are facing students who differ vastly on one of the best predictors of student learning: interest in the subject.

### 2.3.4 Class Size

Armchair theorizing suggests that class size (i.e., the number of enrolled students) ought to be inversely related to SET ratings. Small classes, with 10, 20, or even 30 students, allow each student to have a far greater opportunity to interact with their professors. In contrast, in classes beyond 20 or 30 students, professors are unlikely to learn even student names. Surprisingly, in the first meta-analysis of SET/class size relationship, Feldman (1984) concluded that the average SET/class size correlation was only  $r = -0.09$  (corresponding to  $d = -0.18$ ). Fifteen years later, Aleamoni (1999) summarily declared the notion that class size can affect student ratings to be



a myth. Another 10 years later, Gravestock and Gregor-Greenleaf (2008) concluded that “the correlation between class size and ratings is statistically insignificant and is therefore not viewed as having any impact on validity.”

Our review (Uttl et al., 2018) of over 100 studies that examined the relationship between SET and class size, including those reviewed by Feldman (1984), revealed that the vast majority of these studies did not report sufficient information to interpret their findings. For example, many studies did not report the smallest class size, did not report the largest class size, did not report the number of classes within each class size category, did not examine the linearity of SET/class size relationship, did not examine whether there was a decline in SET for classes with fewer than 20 or 30 students, did not show scatterplots of SET/class size relationships, had very small sample sizes, included extreme outliers, etc.

When only studies that reported sufficient data to plot the relationship between SET and class size and examined, the decline in SET is initially steep and then levels off for class sizes between 30 and 50 students. The overall decline is about 0.5 point on 1–5 rating scale. When each study’s data are standardized using the smallest class size group in each study as a reference group and the average standard deviation of SET means within each study, the declines in SET ratings to class sizes up to 30 or 50 students amount to about 0.5 standard deviation and that the declines continue even thereafter but at a much lower rate. Accordingly, disregarding uninterpretable studies, the evidence clearly shows that declines in SET ratings are steep as class size increases to 30–50 students, and that SET declines level off thereafter.

### **3 SET Are Influenced by Student Preference Factors (SPFs) Whose Consideration Violates Human Rights Legislation**

A substantial body of research has also reported that SET are influenced by factors whose consideration in high-stakes personnel decisions violates human rights legislation such as professors accent, nationality, ethnicity, race, age, gender, etc.

#### **3.1 *Attractiveness/Hotness***

Do students prefer attractive/hot young professors to unattractive/not so hot professors? Using the [www.ratemyprofessor.com](http://www.ratemyprofessor.com) rating data for 6,852 US faculty, Felton et al. (2008) found that Quality (average of Clarity and Helpfulness ratings) was strongly correlated with instructor Hotness ([www.ratemyprofessor.com](http://www.ratemyprofessor.com) discontinued Hotness scale in 2018 in response to a social media campaign against it),  $r = 0.64$ . Hotness was similarly correlated with Helpfulness,  $r = 0.64$ , and Clarity,  $r = 0.60$ , and only moderately correlated with Easiness,  $r = 0.39$ . Accordingly, attractive/hot

professors receive much higher ratings on Clarity, Helpfulness as well as Easiness. One may argue that [www.ratemyprofessor.com](http://www.ratemyprofessor.com) is low quality data, unlike carefully designed SET. However, this argument fails for two reasons: First, [www.ratemyprofessor.com](http://www.ratemyprofessor.com) Overall Quality ratings correlate highly with in class instructor SET ratings with  $r$ s ranging from 0.66 to 0.69 (Coladarci & Kornfield, 2007; Sonntag et al., 2009; Timmerman, 2008). Second, [www.ratemyprofessor.com](http://www.ratemyprofessor.com) ratings are affected by various TEIFs, SPFs, and IDSLFs just as SETs are.

### 3.2 *Accent/Ethnicity/Nationality*

In one of the most extensive studies, Subtirelu (2015) examined the [ratemyprofessor.com](http://www.ratemyprofessor.com) ratings of 2,192 professors with US last names vs. professors with Chinese or Korean last names teaching in the USA. Subtirelu found that professors with US last names received ratings 0.60–0.80 points higher (on 5-point scale) on Clarity and 0.16–0.40 points higher on Helpfulness.

### 3.3 *Gender*

Hundreds of studies have examined gender differences in SET ratings. In general, gender differences in SET ratings are (a) minimal and (b) inconsistent. Moreover, most of the research has compared SET ratings of men vs. women within the university, faculty, or department. However, these studies are impossible to interpret because presence or absence of gender differences does not indicate the presence or absence of gender bias. Gender differences could arise, be reduced, or even masked by a number of different factors including but not limited to gender differences in teaching ability, gender differences in ability to satisfy students, gender differences in courses taught by men vs. women (quantitative vs. non-quantitative, nursing vs. computer science), and gender differences in ability to bake tasty treats for students (see below). However, three recent studies have claimed to show a large bias against female professors and have been widely cited for this claim: Boring (2015, 2017), MacNeill et al. (2015), and Mitchell and Martin (2018). However, a detailed review of these studies does not support their authors' conclusions as detailed below.

Boring (2015, 2017) examined gender differences in SET ratings using a French university's SET ratings of 372 fixed contract instructors teaching seminar sections of introductory courses. Boring found that male teachers received slightly higher ratings than female teachers, mainly because male students rated male teachers somewhat higher than female students (3.20 vs. 3.06 corresponding to approximately 0.2 SD). A re-analysis of Boring's (2015) data set by Boring et al. (2016) shows that the SET/Instructor Gender correlation was only 0.09, corresponding to approximately 0.2 SD. Accordingly, the Boring et al. data suggest that gender differences are small rather than large. However, Boring's data set does not allow the conclusion that

the relatively small differences in SET ratings are evidence of bias against female teachers for at least the following reasons: First, the students were not randomly assigned to seminar sections. For example, students selected whether they took early morning, mid morning, noon, mid afternoon, or late afternoon sections. Second, the students knew the grades given to them by their teachers before they completed SET. Third, there were substantial differences in the experience of female vs. male teachers. For example, a much larger proportion of male teachers had expertise in the field whereas a much larger proportion of female teachers were only PhD students. These experience differences alone could explain the small differences in SET ratings. Fourth, the seminar section teachers were free to teach their section whichever way they liked, used different assignments, etc., and thus, it is impossible to attribute the small differences in ratings to bias.

MacNell et al. (2015) examined the SET ratings of one female and one male instructor of an online course when students were either truthfully told the gender of each instructor (True Gender) or when students were misled about the instructors' genders (and told that each instructor's gender was in fact the opposite of what it was) (False Gender). Both instructors interacted with their students exclusively online, through discussion boards and emails; graded students work at the same time; used the same grading rubrics and co-ordinated their grading to ensure that grading was equitable in their sections. Based on the results of their experiment, MacNell et al. concluded that "Students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender." However, MacNell et al.'s data suffer from several fundamental flaws that render them uninterpretable and MacNell et al.'s conclusions unwarranted (Uttl & Violo, 2021). First, MacNell et al.'s sample of students in each of the four conditions was extremely small, ranging from 8 to 12 students. Second, MacNell et al.'s conclusions depend on three outliers in their small data set—three students who gave their instructors the lowest possible rating on all or nearly all items. When the three outliers are removed from the data set, students rated the actual female instructor numerically higher than the actual male instructor regardless of whether the students were given the actual or false gender of the instructors. Third, MacNell et al.'s study included only one female and one male instructor. It is unwarranted to draw inferences from this small sample size in one study to how students rate female vs. male instructors in general.

Similarly, Mitchell and Martin (2018) examined SET ratings of one female (Mitchell) and one male (Martin) professor teaching different sections of the same online course and found that "a male instructor administering an identical course as a female instructor receives higher ordinal scores in teaching evaluations." Mitchell and Martin argued that their findings were evidence of gender bias as "the only difference in the courses was the identity of the instructor." However, the sections differed or may have differed in many aspects: (a) students' work was graded by different graders whose strictness varied, (b) Drs. Mitchell and Martin held face to face office hours, (c) Drs. Mitchell and Martin may have had different email styles, (c) Mitchell's ratings were based on approximately three times as many responses as

Martin's ratings, (d) Mitchell and Martin may have taught at different times of the day, etc. Moreover, Mitchell and Martin's argument that questions in Instructor/Course, Course, and Technology related to "characteristics that are specific to the course" and do not vary across the sections is simply incorrect. The questions in these categories asked, for example, what "the instructor" did and it ought to be self-evident that different instructors may do things differently, and thus, differences in ratings need not reflect gender bias. Finally, and importantly, just as with MacNell et al. (2015) study, one ought not to make sweeping conclusions about how two categories differ based on differences between two exemplars, one drawn from each of the two categories. This sample size equals one type of research is unlikely to describe what two populations are like.

## **4 SET Are Influenced by Chocolates, Course Easiness, and Other Incentives**

SET ratings are also influenced by numerous factors whose consideration in evaluation of faculty is ill-advised or detrimental to student learning including course difficulty; distribution of chocolates, cookies, and tasty baked goods; and non enforcement of course policies including academic dishonesty and student codes of conduct policies.

### ***4.1 Course Difficulty***

Using data for 3,190 professors from US universities, Felton et al. (2004) found a moderately strong correlation between Quality and Easiness of 0.61. Moreover, the Quality/Easiness relationship became stronger as more ratings were available for each faculty member. Whereas for professors with 10–19 ratings, the Quality/Easiness correlation was 0.61, the correlation reached 0.76 for faculty with 50–59 ratings. The moderate to strong relationship between Quality and Easiness has been subsequently replicated by a number of studies including Felton et al. (2008), Rosen (2018), and Wallisch and Cachia (2019). Wallisch and Cachia (2019) confirmed a steep and accelerated decline of [www.ratemyprofessor.com](http://www.ratemyprofessor.com) Overall Quality ratings (rated on 1–5 point scale) with increasing Course Difficulty (reverse of easiness) (rated on a 1–5 point scale). For each 1.0 point increase in Difficulty, Overall Quality ratings decreased by approximately 0.6 points.

## 4.2 *Chocolates and Cookies*

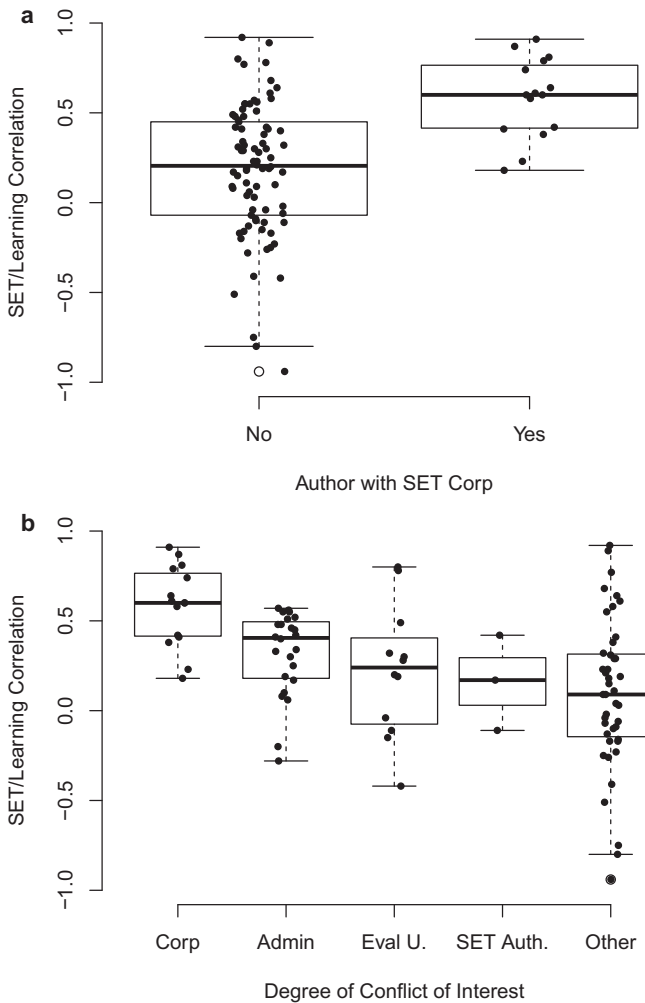
Two randomized studies demonstrate the power of chocolates and cookies in improving SET ratings. In one of the earlier randomized studies, Youmans and Jee (2007) examined whether providing small chocolate bars would result in higher SET ratings in two statistics and one research methods class. Students who were offered chocolate bars rated their instructor substantially higher than students who were not offered chocolate bars ( $d = 0.33$ ). In another randomized study, Hessler et al. (2018) conducted a single-center randomized control group trial to determine whether the availability of chocolate cookies affects SET ratings. Relative to the no-cookie groups, the cookie groups rated teachers as well as the course material much higher,  $d = 0.68$  and  $d = 0.66$ , respectively. Accordingly, at minimum, chocolates and chocolate cookies are both very effective ways to increase one's SET ratings.

## 5 SET Findings Vary with Conflict of Interest

Uttl et al. (2019) have recently shown that the correlations between SET and learning/achievement in the multisection studies discussed above depend not only on their sample size but also on their authors' degree of conflict of interest (perceived or actual). Figure 3, Panel A shows that authors with SET corporations (Corp) reported much higher SET/learning correlations than authors with no such ties,  $r = 0.58$  vs.  $r = 0.18$ , respectively. However, as shown in Panel B, conflict of interest is not limited to authors with direct financial gains from selling SET but also extends to authors with other non financial conflicts of interest such as administrative (Admin) and evaluation units (Eval U) ties. These findings are particularly troubling; they suggest that in addition to the poor methodology employed by many SET studies (e.g., small sample sizes, insufficient method descriptions, failure to consider outliers), many SET research findings may also be the result of their authors financial and other interests, whether these biases were conscious or unconscious.

## 6 Discussion

SET do not measure teaching effectiveness and students do not learn more from more highly rated professors. Until recently, meta-analyses of multisection studies have been cited as the best evidence of SET validity. Those meta-analyses, however, were fundamentally flawed. The re-analyses of the previous meta-analyses as well as the new updated meta-analyses of multisection studies show that SET are unrelated to student learning in multisection designs. Accordingly, SET ought not to be used to measure faculty's teaching effectiveness.



**Fig. 3** SET/learning correlations and conflict of interest. Panel **A** shows that authors with SET corporations (Corp) reported much higher SET/learning correlations than authors with no such corporate ties. Panel **B** shows that authors with other conflicts of interests including administrative (Admin) and evaluation units (Eval U) also reported higher SET/learning correlations whereas authors with no identifiable conflicts of interest reported near zero SET/learning correlations (the figures are adapted from Uttl et al. 2019)

Regardless of what SET actually measure, SET are substantially influenced by (1) numerous factors not attributable to professors, including students' intelligence and prior knowledge, students' motivation and interest, class size, and course subject; (2) factors attributable to professors but whose consideration in high-stakes personnel decisions violates human rights legislation such as accent, race, ethnicity, national origin, age, and hotness/sexiness; and (3) factors attributable to professors but whose

consideration is at minimum unwise and/or detrimental to student learning, including course difficulty and availability of chocolates and cookies.

Although some SET systems attempt to adjust for influences of various TEIFs, SPFs, and DSLFs, these attempts are ultimately futile because no SET system can nor does adjust for all demonstrated effects of TEIFs, SPFs, and DSLFs, nor for effects of possible TEIFs, SPFs, and DSLFs. Even adjusting only for the factors reviewed above would likely be impossible. For example, to adjust for factors attributable to students, one would have to administer highly reliable and valid tests of student intelligence, prior knowledge, motivation, interest, racism, accent preference, hotness preferences, etc., then calculate average class intelligence, prior knowledge, motivation, interest, racism, accent preferences, hotness preferences, etc., and then develop some adjustment system. No one has done it so far and no one is likely to do so in the foreseeable future.

SET measure student satisfaction, that is, “a fulfillment of need or want” or “a happy or pleased feeling because of something that you did or something that happened to you” ([www.m-w.com](http://www.m-w.com)). One may argue that student satisfaction is important and that student satisfaction is properly used or ought to be used in high-stakes personnel decisions such as hiring, firing, promotion, merit pay, and teaching awards. However, the fundamental problem with using student satisfaction at all to evaluate faculty is that it depends on factors not attributable to professors.

Moreover, making high-stakes personnel decisions by comparing faculty’s SET ratings to university, faculty, or departmental norms, sets up and fuels a race among faculty members to beat at least 30, 50 or, 70% of their colleagues depending on the particular norm-referenced criteria for unsatisfactory, “in need of improvement”, etc. adopted by their institution. This race for higher and higher SET ratings is what a number of writers believe is the principal cause of run-away grade inflation and work deflation (Crumbly & Reichelt, 2009; Emery et al., 2003; Haskell, 1997; Stroebe, 2016, 2020). Although SET were relatively rare prior to 1970s, today they are used by almost all colleges and universities in North America and in many other countries to evaluate teaching effectiveness (Seldin, 1993). Accordingly, the race for higher SET pressures faculty to satisfy their students’ needs and wants, in particular, to increase grades, reduce workload, tolerate academic dishonesty, avoid topics that may antagonize some students, etc. Indeed, the grades have been increasing and have risen from C grades being the most frequently awarded grades in 1970s to A grades being the most frequently awarded grades today (Rojstaczer & Healy, 2010, 2012). At the same time, students report spending less and less time on their studies (Fosnacht et al., 2018; Rojstaczer & Healy, 2010). Whereas in the 1960s students in US were spending on average about 2 h studying outside of the class for each hour in the class, today students are spending only about 1 h. These two trends are nothing short of astonishing when one considers that the average intelligence and ability of students entering colleges and universities has declined over the last 50–100 years, as the proportions of high school graduates entering universities and colleges has increased from approximately 5% to more than 50% or even 70% depending on the country, state, and province (US Census Bureau, 2019). Notably, SET are not the only cause of grade inflation and work deflation. Other related causes include

colleges and universities' focus on high student retention; pressure on professors to limit percentages of D, F, and W (withdrawal) grades (explicitly requiring professors to increase grades); and business culture that not only strives for happy customers whose needs and wants need to be satisfied but also for as many customers as possible.

In the first public legal case of its kind, Ryerson University was forbidden from using SET as a measure of teaching effectiveness (*Ryerson University v. Ryerson Faculty Association*, 2018 CanLII 58,446, available at [www.canlii.org](http://www.canlii.org)). The arbitrator Williams stated:

That evidence, as earlier noted, was virtually uncontradicted. It establishes, with little ambiguity, that a key tool in assessing teaching effectiveness is flawed, while the use of averages is fundamentally and irreparably flawed. It bears repeating: the expert evidence called by the Association was not challenged in any legally or factually significant way. As set out above, the assessment of teaching effectiveness is critical, for faculty and the University, and it has to be done right. The ubiquity of the [SET] tool is not a justification, in light of the evidence about its potential impact, for its continuation, or for mere tinkering.

The SET ratings also run afoul of at least some codes of ethics. For example, Canadian Code of Ethics for Psychologists (Canadian Psychological Association, 2017) makes it clear that psychologist not only has a duty to not participate in incompetent and unethical behavior, such as evaluating their colleagues using invalid and biased SET tools, they also have a responsibility to call out "incompetent and unethical behavior, including misinterpretations or misuses of psychological knowledge and techniques" (Ethical Standard IV.13).

Notwithstanding the above criticisms, student surveys may continue to be useful for formative uses, that is, for improving instruction when professors themselves design or select questions relevant to their teaching methods and courses, and when SET are provided only to professors themselves to ensure that they are not misused, not used for summative uses, and used only for formative uses or to raise alarm about some ineffective teaching behaviors (e.g., not showing up for one's classes).

Finally, and importantly, this review of SET research highlights the need for transparent, replicable, and methodologically strong research, conducted by researchers with no conflict of interest and no interest in particular findings. The SET literature is replete with unsubstantiated and contradictory findings based on poor methods. As detailed above, Cohen's (1981) widely cited evidence of SET validity turned out to be an artifact of poor methods and failure to take into account small sample bias and students' prior ability. Similarly, Feldman's (1984) finding of minimal effect of class size and Aleamoni's (1999) later dismissal of the idea that class size is related to SET ratings as a myth were similarly based on poor methods and failure to adequately review the previous findings. And MacNeill et al. (2015) claim of gender bias against women hinges in its entirety on three outliers, three students who disliked their instructors so much as to give them the lowest possible rating on all or nearly all items. Significantly, as shown by Uttl et al. (2019), the reported findings may be greatly influenced by a conflict of interest. It is clear that any review of this literature needs to be approached with an attitude of a detective rather than simply accepting what is written in studies' abstracts in order to ferret out true findings supported by evidence from uninterpretable and unwarranted claims.



In conclusion, continued use of SET in high-stakes personnel decision such as hiring, firing, promotion, merit pay, and teaching award is not evidence based. The evidence is that (a) students do not learn more from more highly rated professors; (b) SET are biased by a variety of factors not attributable to professors; (c) SET run afoul to human rights legislation, and (d) SET are easily manipulated by small chocolates such as Hershey's kisses, course easiness, and other factors. In short, SET do not measure faculty's teaching effectiveness and their use in high-stakes personnel decisions is improper, unethical, and ought to be discontinued immediately.

## References

- Abrami, P. C., & d'Apollonia, S. (1999). Current concerns are past concerns. *American Psychologist*, 54(7), 519–520. <https://doi.org/10.1037/0003-066X.54.7.519>.
- Aleamoni, L. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, 13(2), 153–166. <https://doi.org/10.1023/A:1008168421283>.
- Beran, T., & Violato, C. (2009). Student ratings of teaching effectiveness: Student engagement and course characteristics. *Canadian Journal of Higher Education*, 39(1), 1–13.
- Boring, A. (2015). *Gender Biases in student evaluations of teachers* (No. 2015–13). Documents de Travail de l'OFCE. Observatoire Francais des Conjonctures Economiques (OFCE). <https://ideas.repec.org/p/fce/doctr/1513.html>. Accessed 4 June 2020.
- Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145, 27–41. <https://doi.org/10.1016/j.jpube.2016.11.006>.
- Boring, A., Ottononi, K., & Stark, P. B. (2016). *Student evaluations of teaching are not only unreliable, they are significantly biased against female instructors*. The London School of Economics and Political Science. <https://doi.org/10.14293/s2199-1006.1.sor-edu.aetbzc.v1>.
- Canadian Psychological Association. (2017). *Canadian code of ethics for psychologists* (4th ed.). Canadian Psychological Association.
- Centra, J. A. (2009). *Differences in responses to the Student Instructional Report: Is it bias?* Educational Testing Service.
- Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, 51(3), 281–309. <https://doi.org/10.3102/00346543051003281>.
- Coladarsi, T., & Kornfield, I. (2007). RateMyProfessors.com versus formal in-class student evaluations of teaching. *Practical Assessment, Research & Evaluation*, 12(6), 1–15.
- Crumbley, D. L., & Reichelt, K. J. (2009). Teaching effectiveness, impression management, and dysfunctional behavior: Student evaluation of teaching control data. *Quality Assurance in Education: An International Perspective*, 17(4), 377–392. <https://doi.org/10.1108/09684880910992340>.
- Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, 11(1), 37–46. <https://doi.org/10.1108/09684880310462074>.
- Feldman, K. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education*, 21(1), 45–116. <https://doi.org/10.1007/BF00975035>.
- Felton, J., Koper, P. T., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on Ratemyprofessors.com. *Assessment & Evaluation in Higher Education*, 33(1), 45–61. <https://doi.org/10.1080/02602930601122803>.
- Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: The relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education*, 29(1), 91–108. <https://doi.org/10.1080/0260293032000158180>.

- Fosnacht, K., McCormick, A. C., & Lerma, R. (2018). First-year students' time use in college: A latent profile analysis. *Research in Higher Education*, 59(7), 958–978. <https://doi.org/10.1007/s11162-018-9497-z>.
- Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends*. Higher Education Quality Council of Ontario. <https://deslibris.ca/ID/215362>. Accessed 22 Feb 2020.
- Haskell, R. E. (1997). Academic freedom, tenure, and student evaluation of faculty. *Education Policy Analysis Archives*, 5, 6. <https://doi.org/10.14507/epaa.v5n6.1997>.
- Hessler, M., Pöpping, D. M., Hollstein, H., Ohlenburg, H., Arnemann, P. H., Massoth, C., et al. (2018). Availability of cookies during an academic course session affects evaluation of teaching. *Medical Education*, 52(10), 1064–1072. <https://doi.org/10.1111/medu.13627>.
- Hoyt, D. P., & Lee, E. (2002). *Technical Report No. 12: Basic data for the revised IDEA system*. The IDEA Center.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>.
- MacNell, L., Driscoll, A., & Hunt, A. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4), 291–303. <https://doi.org/10.1007/s10755-014-9313-4>.
- Marsh, H. W. (1980). *Students' evaluations of college/university teaching: A description of research and an instrument*.
- Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, 83(2), 285–296. <https://doi.org/10.1037/0022-0666.83.2.285>.
- Mitchell, K. M. W., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, 51(3), 648–652. <https://doi.org/10.1017/S104909651800001X>.
- Orpwood, G., & Brown, E. S. (2015). Closing the numeracy gap. CGC Educational Communications. [http://www.numeracygap.ca/assets/img/Closing\\_the\\_numeracy\\_Executive\\_Summary.pdf](http://www.numeracygap.ca/assets/img/Closing_the_numeracy_Executive_Summary.pdf). Accessed 20 May 2020.
- Rojstaczer, S., & Healy, C. (2010). Grading in American colleges and universities. *Teachers College Record*.
- Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940–2009. *Teachers College Record*, 114(7), 23.
- Rosen, A. S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: A large-scale study of RateMyProfessors.com data. *Assessment & Evaluation in Higher Education*, 43(1), 31–44. <https://doi.org/10.1080/02602938.2016.1276155>.
- Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education; Washington*, 39(46), A40.
- Sonntag, M. E., Bassett, J. F., & Snyder, T. (2009). An empirical test of the validity of student evaluations of teaching made on RateMyProfessors.com. *Assessment & Evaluation in Higher Education*, 34(5), 499–504. <https://doi.org/10.1080/02602930802079463>.
- Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 11(6), 800–816. <https://doi.org/10.1177/17456916166650284>.
- Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology*, 42(4), 276–294. <https://doi.org/10.1080/01973533.2020.1756817>.
- Subtirelu, N. C. (2015). “She does have an accent but...”: Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society*, 44(1), 35–62. <https://doi.org/10.1017/S0047404514000736>.

- Timmerman, T. (2008). On the validity of RateMyProfessors.com. *Journal of Education for Business*, 84(1), 55–61. <https://doi.org/10.3200/JOEB.84.1.55-61>.
- US Census Bureau. (2019). CPS historical time series tables: Table A-1 Years of school completed by people 25 years and over, by Age and Sex: Selected years 1940 to 2019. US Census Bureau. <https://www.census.gov/data/tables/time-series/demo/educational-attainment/cps-historical-time-series.html>. Accessed 20 May 2020.
- Uttl, B., Bell, S., & Banks, K. (2018). Student evaluation of teaching (SET) ratings depend on the class size: A systematic review (No. 8110392). In *Proceedings of International Academic Conferences*. International Institute of Social and Economic Sciences. <https://ideas.repec.org/p/sek/iacpro/8110392.html>. Accessed 27 May 2020.
- Uttl, B., Cnudde, K., & White, C. A. (2019). Conflict of interest explains the size of student evaluation of teaching and learning correlations in multisection studies: A meta-analysis. *PeerJ*, 7(7), e7225. <https://doi.org/10.7717/peerj.7225>.
- Uttl, B., & Kibreab, M. (2011). Self-report measures of prospective memory are reliable but not valid. *Canadian Journal of Experimental Psychology = Revue Canadienne De Psychologie Experimentale*, 65(1), 57–68. <https://doi.org/10.1037/a0022843>.
- Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. *PeerJ*, 5(5), e3299. <https://doi.org/10.7717/peerj.3299>.
- Uttl, B., & Violo, V. (2021). Small samples, unreasonable generalizations, and outliers: Gender bias in student evaluation of teaching or three unhappy students? *ScienceOpen Research*. <https://doi.org/10.14293/S2199-1006.1.SOR-2021.0001.v1>.
- Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54, 22–42. <https://doi.org/10.1016/j.stueduc.2016.08.007>.
- Uttl, B., White, C. A., & Morin, A. (2013). The numbers tell it all: Students don't like numbers! *PLoS ONE*, 8(12), e83443. <https://doi.org/10.1371/journal.pone.0083443>.
- Wallisch, P., & Cachia, J. (2019). *Determinants of perceived teaching quality: The role of divergent interpretations of expectations*. <https://doi.org/10.31234/osf.io/dsvvgq>.
- Williams, P. G., Rau, H. K., Suchy, Y., Thorgusen, S. R., & Smith, T. W. (2017). On the validity of self-report assessment of cognitive abilities: Attentional control scale associations with cognitive performance, emotional adjustment, and personality. *Psychological Assessment*. <https://doi.org/10.1037/pas0000361>.
- Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course. *Teaching of Psychology*, 34(4), 245–247. <https://doi.org/10.1080/00986280701700318>.

**Bob Uttl** is a Professor of Psychology at Mount Royal University, Calgary (Canada). Previously he held a variety of academic posts around the world including: Red Deer College, Canada; Tamagawa University, Japan; University of Tsukuba, Japan; Oregon State University, USA; Henry M. Jackson Foundation for Advancement of Military Medicine, USA; and National Institute of Health, USA. His research includes: the relationship between perception, processing resources, and prospective and retrospective memory; changes in cognitive functions due to normal and pathological aging; measurement and research methods; and evaluation of teaching in higher education.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

