On Students’ (Mis)judgments of Learning and Teaching Effectiveness

Shana K. Carpenter∗ and Amber E. Witherby
Iowa State University, USA

Sarah K. Tauber
Texas Christian University, USA

Students’ judgments of their own learning are often misled by intuitive yet false ideas about how people learn. In educational settings, learning experiences that minimize effort and increase the appearance of fluency, engagement, and enthusiasm often inflate students’ estimates of their own learning, but do not always enhance their actual learning. We review the research on these “illusions of learning,” how they can mislead students’ evaluations of the effectiveness of their instructors, and how students’ evaluations of teaching effectiveness can be biased by factors unrelated to teaching. We argue that the heavy reliance on student evaluations of teaching in decisions about faculty hiring and promotion might encourage teaching practices that boost students’ subjective ratings of teaching effectiveness, but do not enhance—and may even undermine—students’ learning and their development of metacognitive skills.

General Audience Summary
As the changing landscape of education provides more freedom and flexibility in the options available to students, it is becoming increasingly important that students be able to successfully evaluate and manage their own learning. This is easier said than done, however, because students often misjudge their own learning of a given topic to be better than it actually is. This common tendency toward overconfidence can be further bolstered by a number of intuitive but misleading factors that enhance students’ subjective impressions of how much they have learned, without always enhancing their actual learning. Students believe, for example, that they learn best from enthusiastic and engaging instructors who provide smooth and well-polished lectures that do not require active class participation. Such factors, although they readily inflate students’ judgments of their own learning, do not consistently enhance students’ actual learning. They also inflate students’ evaluations of the effectiveness of their instructors. Indeed, students’ evaluations of teaching effectiveness can be poor predictors of their actual learning in their courses, and these evaluations can be biased by external factors unrelated to student learning, such as an instructor’s gender, age, attractiveness, and grading leniency. Given the heavy reliance on student evaluations of teaching effectiveness in decisions regarding faculty hiring and promotion, faculty may be incentivized to adopt teaching approaches that boost their evaluations but do not enhance—and could even undermine—students’ academic success.

Keywords: Learning, Metacognition, Education, Illusions of learning, Teaching evaluations

Author Note
Shana K. Carpenter and Amber E. Witherby, Department of Psychology, Iowa State University, USA.
Sarah K. Tauber, Department of Psychology, Texas Christian University, USA.

This material is based upon work supported by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition, Collaborative Grant No. 220020483.
∗ Correspondence concerning this article should be addressed to Shana K. Carpenter, Department of Psychology, Iowa State University, W112 Lagomarcino Hall, 901 Stange Road, Ames, IA 50011, USA. Contact: shacarp@iastate.edu.
Educational practices have changed considerably in recent decades. Due largely to advances in technology, the increasing popularity of non-conventional pedagogical approaches such as flipped classrooms, problem-based learning, and online courses reflect a modern classroom in which the highly structured environment of the past has been replaced by one that allows more freedom and flexibility. Given that students have more choices now than ever before with respect to how and when they learn something, the ability to evaluate one’s own learning might be among the most important skills in 21st century education.

Unfortunately, this skill remains underdeveloped among students. There is often a great divide between what students think they have learned and the measurable evidence of their learning. This manifests through numerous visits to office hours following the first exam in a course, when many students express surprise and confusion upon finding that they have performed lower than expected despite having a strong sense that they understood the material. Beyond an uncomfortable conversation with one’s professor, students’ disappointment in their own performance can have farther-reaching consequences, such as whether or not they decide to persist in the course or stay in college (e.g., Geisinger & Raman, 2013), and how favorably they rate their professor’s teaching skills on the end-of-term course evaluations (e.g., Stroebel, 2016).

Worse yet, these important decisions can be based on factors that do not accurately reflect students’ learning. Decades of research on metacognition have shown that students are not particularly good at evaluating their own learning, and they hold many false assumptions about how people learn. Students have a strong tendency to prefer instructional approaches that enhance their subjective impressions of learning, but that have been shown through empirical research to be ineffective or even counter productive for learning. Given that many colleges and universities rely heavily on students’ subjective evaluations as a measure of teaching effectiveness, the question arises as to whether this system might encourage suboptimal teaching practices that inflate subjective impressions but do not enhance student learning. In this article, we review some of the factors present in educational settings that can mislead students’ judgments of learning and teaching effectiveness, how these judgments relate to students’ actual learning in their courses, and what this means for using student evaluations of teaching as the primary measure for assessing the quality of instruction.

Illusions of Learning: Factors that Inflate Students’ Impressions of Learning but do not Enhance Actual Learning

The most common metacognitive problem is overconfidence. When given some material to learn and asked to estimate how well they will perform on a test over that material, students’ subjective estimates of their own learning often exceed their objective performance. In addition to well-controlled laboratory experiments that have demonstrated this overconfidence bias for a number of years (for a review, see Finn & Tauber, 2015), classroom data also show that students typically expect to perform much better on exams than they actually do (Hacker, Bol, Horgan, & Rakow, 2000; Hartwig & Dunlosky, 2017; Miller & Geraci, 2011), even after having completed multiple exams in a course (Foster, Was, Dunlosky, & Isaacson, 2017).

Overconfidence appears to be a naturally occurring tendency that starts early in development. Even very young children substantially overestimate their own memory abilities. For example, Flavell, Friedrichs, and Hoyt (1970) found that 64% of kindergarten children predicted that they would score perfectly on a fairly difficult picture memory task, for which children typically only recalled about half of the pictures. Children may overestimate their learning because they rely on motivational factors such as wishful thinking (e.g., Scheider, 1998). Although children’s overconfidence is ubiquitous, they can make accurate predictions under some circumstances (Cunningham & Weaver, 1989; Finn & Metcalfe, 2014; Lipko, Dunlosky, Lipowski, & Merriman, 2012; Lipko, Dunlosky, & Merriman, 2009; Lipowski, Merriman, & Dunlosky, 2013; Shin, Bjorklund, & Beck, 2007; Yussen & Berman, 1981; Yussen & Levy, 1975). Slightly older elementary school children also exhibit overconfidence in their memory abilities, and these tendencies persist throughout middle and high school as well (for a review, see Schneider & Löfler, 2016).

These strong tendencies toward overconfidence are hard to overcome. Even students entering college, who have been learning academic material for many years, demonstrate overconfidence in their learning. Further, these tendencies can be bolstered by a number of factors that are widespread in educational settings. So common and intuitive are these factors that many instructors have faith in them as well, and they would readily incorporate them into their teaching based on the logical and strong assumption that these things are good for student learning.

Consider, for example, the very reasonable idea that a lesson should be well-organized. Conscientious instructors spend a great deal of time preparing lessons to provide a logical and organized flow of information, and they practice their lectures to ensure a smooth delivery with the goal that the information should make sense and should not be confusing to students. Likely for these reasons, many handbooks on effective teaching encourage instructors to prepare lectures that are highly organized (Brown & Atkins, 1990; Brown & Race, 2002; Davis, 1993; Ekeler, 1994; Hogan, 1999; Lowman, 1995; Morton, 2009). Indeed, when the content of a lecture is made more organized—for example, by including clarifying statements and transitions between concepts—students perceive the lecture as clear and they also perform better on tests of their knowledge over that lecture (Titsworth, 2001; see also Titsworth & Kiewra, 2004).

Perceptions of clarity and organization do not always coincide with better learning, however. In particular, instructors
who are perceived as more organized—independent of the content they teach—have been shown to increase students’ subjective impressions of how much they have learned, but have no consistent effect on students’ actual learning. This “illusion of learning” has been demonstrated in a series of laboratory-based studies on instructor presentation style (Carpenter, Mickes, Rahman, & Fernandez, 2016; Carpenter, Northern, Tauber, & Tofness, in press; Carpenter, Wilford, Kornell, & Mullaney, 2013; Tofness et al., 2018). In these studies, students viewed a video of an instructor presenting a lecture in a fluent style—standing upright, facing the camera, using vocal inflections, and appropriate gestures—or in a disfluent style—hunching over a podium, reading from notes, speaking in a monotone voice. Afterward, they estimated how much they thought they had learned from the lecture, and then completed a memory test over the lecture content. Everything about the two videos was identical—the same instructor presented the fluent and disfluent lecture and spoke the same scripted content—such that only the presentation style (fluent vs. disfluent) differed. Students who watched the fluent instructor rated the instructor as significantly more organized compared to students who watched the disfluent instructor, and they also judged their own learning to be higher. The higher confidence in learning was merely an illusion, however, as the test scores between the two groups were not significantly different.

Similar illusions of learning can occur for visual aids. Diagrams, pictures, and illustrations are recommended in handbooks on teaching as a means of clearly demonstrating and explaining information (Brown & Atkins, 1990; Orlich, Harder, Callahan, Trevisan, & Brown, 2010). There is evidence that such tools can enhance learning when they provide additional explanatory information that is relevant to the text (Carney & Levin, 2002; Levie & Lentz, 1982). However, students have a tendency to over-endorse the effectiveness of pictures and illustrations, even when these images are simply decorative and do not enhance their understanding of what they are reading. The mere presence of images or photographs in text material, for example, does not consistently enhance learning of that material, but significantly increases students’ confidence that they have learned it (Carpenter & Olson, 2012; Lenzner, Schnotz, & Mueller, 2013; Serra & Dunlosky, 2010). Along similar lines, the presence of multimedia animations within a lesson can enhance students’ learning in some circumstances (Berney & Bétrancourt, 2016), but even when it does not, students tend to be confident that it does (Paik & Schraw, 2013).

Thus, the organization of a lesson and the presence of visual aids seem intuitively advantageous for learning and can indeed enhance learning under some circumstances. The finding that these things can increase students’ impressions of their learning without always increasing actual learning, however, suggests that the factors responsible for effective learning are not always consistent with pre-existing assumptions. To the contrary, the appearance of clarty, organization, and visual representations can sometimes mislead students into thinking they have learned more than they actually have.

Illusions of Learning and Student Evaluations of Instructors

Just as students’ subjective impressions of their own learning can be misled, so can their impressions of the effectiveness of their instructors. In the studies on instructor fluency (Carpenter et al., 2016, in press, 2013; Tofness et al., 2018), students were asked to rate the instructor on overall effectiveness as well as the instructor’s level of organization, preparation, and knowledge. In all studies, the fluent instructor received substantially higher ratings on all of these measures than did the disfluent instructor. The invariant effects of instructor fluency on actual learning, however, indicate that students’ impressions of effective teaching did not coincide with effective learning.

Enthusiasm may be another false indicator of teaching effectiveness. Nobody can blame students for preferring a lively and enthusiastic instructor over a boring one. Indeed, enthusiastic instructors readily increase students’ affective responses such as self-reported ratings of enjoyment (Frenzel, Goetz, Luedtke, Pekrun, & Sutton, 2009), interest (Keller, Goetz, Becker, Morger, & Hensley, 2014), and engagement (Zhang, 2014). As well, instructor enthusiasm is commonly considered a quality of effective teaching (e.g., Minor, Onwuegbuzie, Witcher, & James, 2002). However, there is a lack of evidence regarding whether instructor enthusiasm has positive and consistent effects on students’ actual learning. Laboratory-based studies that manipulate instructor enthusiasm—through brief video-recorded lectures (ranging from 5 to 30 min) of an instructor delivering a lesson in an enthusiastic style (engaging behavior, humor, and personal anecdotes) versus the same instructor delivering the same lesson without these attributes—show that instructor enthusiasm inflates student evaluations of teaching effectiveness, but does not reliably affect students’ test scores over the content (Meier & Feldhusen, 1979; Motz, de Leeuw, Carvalho, Liang, & Goldstone, 2017; Perry, Abrami, & Leventhal, 1979; Williams & Ware, 1976).

Williams and Ceci (1997) found the same pattern of results in a real course in which an experienced instructor taught the course in his usual style, and then the following term deliberately changed his presentation style to be more enthusiastic while keeping all other aspects of the course as identical as possible. On the end of term course evaluations, students in the “enthusiastic” class rated the instructor as significantly more effective than did students who took the course the previous term. Compared to students who took the course the previous term, students in the enthusiastic class rated the instructor higher on various instructor attributes—such as organization and level of knowledge—and also rated the course higher on aspects that were identical between the two terms, such as the quality of the textbook. Furthermore, students who had the enthusiastic instructor estimated that they had learned more
than students who took the course the previous term, when in fact the course grades between the two groups were nearly identical.

In an experimental field study, Bettencourt, Gillett, Gall, and Hull (1983) randomly assigned math instructors to a training program designed to increase their enthusiasm in the classroom, or to a control group that did not complete this training. Although this program had the intended effect of increasing the instructors’ displays of enthusiastic behaviors while teaching, such behaviors did not enhance students’ learning. That is, the test scores of students who were taught by the trained instructors were no different from those of students who were taught by the control group of instructors.

Correlational data from large courses also show that the impression of an engaging, enthusiastic, or fluent instructor coincides with students’ judgments of teaching effectiveness but not actual learning. Serra and Magreehan (2016) found that in a large introductory psychology course, students’ evaluations of their instructor and estimates of how much they had learned in the course correlated positively with their ratings of several instructor-based attributes that reflected fluency and engagement—such as the instructor’s clarity of speech, ability to maintain students’ attention, and use of visual aids—even after controlling for students’ final grades in the course.

The “Dr. Fox effect” has become a classic example of the persuasive but misleading power of instructor enthusiasm. During a teacher training conference, Naftulin, Ware, and Donnelly (1973) hosted a lecture on educational applications of mathematical game theory, to be delivered by Dr. Myron L. Fox, an expert on mathematics applied to human behavior. “Dr. Fox” really knew nothing about the topic and was in fact a Hollywood actor who had been hired and asked to prepare the lecture from a brief article in a popular science magazine. He had further been instructed to make the lecture intentionally meaningless by including vague material, contradictory statements, redundancies, and humorous content that was unrelated to the lecture topic. Despite the nonsense material, Dr. Fox delivered the lecture in a highly enthusiastic and passionate style. Afterward, feedback solicited from the audience members indicated an overwhelmingly positive reaction, with over 90% of attendees reporting that the lecture was well-organized, interesting, and contained clear examples. This now-classic demonstration echoes the findings from empirical research showing that desirable instructor behaviors such as fluency and enthusiasm can be powerful contributors to the positive impressions formed about those instructors, regardless of the content that is taught or learned.

Beyond the personal characteristics of instructors, the pedagogical methods they use can also mislead students’ judgments of teaching effectiveness. “Active learning” has become a polarizing concept that is often embraced by faculty but strongly resisted by students. Compared to the traditional lecture approach during which students sit and listen to an instructor, many disciplines are starting to incorporate more student involvement in the form of hands-on activities during class, peer interaction, and pre-class learning activities. The passive lecture gives the impression of a fluent, smooth, and seamless learning experience, whereas active learning creates a more disjointed, less fluent experience, in that students may need to think more deeply about, and struggle with, the material to understand and apply it. It is perhaps no surprise, therefore, that many students resist active learning techniques on the grounds that they feel they are not learning (e.g., Seidel & Tanner, 2013).

This impression again may be an illusion, possibly driven by similar factors that underlie the illusion of learning due to instructor fluency. In one recent study (Deslauriers, McCarty, Miller, Callaghan, & Kestin, 2019), introductory physics students were randomly assigned to experience a class lesson over the same material that involved either passive lecture (i.e., the instructor provided the solutions to all practice problems, with no student interaction) or active learning (i.e., the students attempted the problems first in small groups, followed by the instructor presenting the solutions). Following the lesson, students rated how much they felt they had learned and how effective they felt the instructor was, and then completed a multiple-choice test over the material from the lesson. Students who experienced the passive lecture gave significantly higher ratings of their own learning, and they also rated the instructor as significantly more effective, than did students who experienced the same lesson via active learning. Scores on the test at the end of the lesson, however, revealed a significant advantage for students who experienced active learning compared to students who experienced the passive lecture. Although more research is needed on active learning and whether its benefits are specific to particular courses and subject matter—including the effects of mixing active learning and lecture-based approaches—these results show that under conditions when active learning is effective, students are not aware of these benefits.

These results highlight an important disconnect between students’ impressions of effective teaching and the actual evidence of it. Students routinely associate “effective” teaching with experiences that feel easy, smooth, fluent, or enjoyable. As the evidence shows, however, such methods do not always promote learning and could even undermine it. If curricular decisions were made on the basis of students’ subjective impressions of teaching effectiveness, the study by Deslauriers et al. (2019) suggests that this decision would result in the implementation of inferior pedagogical approaches. Indeed, in this study students strongly endorsed a teaching method that resulted in an approximately 10-point disadvantage (a letter grade drop) in their test scores. This raises serious questions about the validity of student evaluations of teaching effectiveness.

### Student Evaluations of Teaching and their Relation to Learning Outcomes

Student evaluations have long been utilized in colleges and universities as a means of providing feedback to instructors about meeting curricular goals. The earliest research on student evaluations dates back to the 1920s (e.g., Remmers & Brandenburg, 1927). Ideally, these evaluations should provide a measure of how effective a given instructor is at promoting student learning. The evidence reviewed above, however, suggests that student evaluations may not be reliable indicators of
learning and could be influenced by a number of factors that are unrelated to learning.

This could explain the inconsistent relationship between students’ evaluations of teaching effectiveness and student learning outcomes in real courses. Data collected from a variety of courses have revealed every possible outcome, in that this relationship is sometimes positive (e.g., Bryson, 1974; Cohen, 1981; Sullivan & Skanes, 1974), sometimes negative (e.g., Braga, Paccagnella, & Pellizzari, 2014; Carrell & West, 2010; Yunker & Yunker, 2003), and sometimes non-existent (e.g., Boring, Ottoboni, & Stark, 2016; Palmer, Carliner, & Romer, 1978; Uttl, White, & Gonzalez, 2017).

Interestingly, the direction of the relationship between student evaluations of teaching effectiveness and student learning depends on how learning is measured. One way to operationalize learning is via grades students receive in the courses taught by the instructors who are being evaluated. Hundreds of studies have explored how this measure of learning relates to student evaluations of teaching effectiveness, and numerous reviews and meta-analyses have been published cumulating their results (e.g., Clayson, 2009; Cohen, 1981; Feldman, 1976; Marsh, 2007b; McCallum, 1984; Uttl et al., 2017). Although some researchers have found no relationship between student evaluations of teaching effectiveness and the grades that students earn (e.g., Boring et al., 2016; Doyle & Whitey, 1974), the bulk of this research has concluded that there is a weak-to-moderate positive relationship (e.g., Brockx, Spooren, & Mortelmans, 2011; Centra, 1977; Frey, 1976; Remedios & Lieberman, 2008; for reviews, see Cohen, 1981; Feldman, 1976). Thus, students tend to perform better in the courses taught by the instructors they have rated as more effective.

Interpreting this positive correlation presents a chicken-egg problem, however. It is not clear whether students performed well because the instructor was effective, or whether the instructor was rated as effective because the students performed well. The positive relationship between students' evaluations of teaching effectiveness and their grades in the course could occur, for example, due to grading leniency and the tendency for students to “reward” an instructor who gives good grades (for these and other issues, see Gillmore & Greenwald, 1999; Greenwald & Gillmore, 1997a; Marsh, Fleiner, & Thomas, 1975).

In response to this issue, researchers have explored whether student evaluations of teaching predict student performance on measures of learning that are less susceptible to subjective influences like grading leniency. The outcomes of this research are less consistent. When considering the relationship between student evaluations of teaching and student performance on standardized exams (i.e., multiple instructors use the same common exam for a given class), some studies show a positive relationship (e.g., Beleche, Fairris, & Marks, 2012; Bryson, 1974; Sullivan & Skanes, 1974), some show a negative relationship (e.g., Rodin & Rodin, 1972), and some show no relationship (e.g., Braskamp, Caulley, & Costin, 1979; Endo & Della-Piana, 1976; Greenwood, Hazelton, Smith, & Ware, 1976). In addition, in a recent review, Uttl et al. (2017) argued that the positive relationships that were consistently found by using class grades and exam performance as measures of achievement were obscured by small sample sizes and publication bias. In their updated meta-analysis, Uttl et al. found no evidence of a relationship between student evaluations of teaching effectiveness and student achievement on both standardized and non-standardized assessments.

Perhaps the most powerful evidence of effective teaching is durable, long-lasting learning. If an instructor accomplishes this goal, there should be a positive relationship between an instructor’s measurable effectiveness and students’ performance in follow-up courses that are relevant to what they learned from that instructor. A small but growing number of studies has investigated this relationship. All the available evidence has demonstrated a negative correlation between students’ evaluations of teaching and their performance in follow-up courses (Braga et al., 2014; Carrell & West, 2010; Johnson, 2003; Weinberg, Fleisher, & Hashimoto, 2007; Yunker & Yunker, 2003; for a review, see Kornell & Hausman, 2016). To illustrate, Yunker and Yunker (2003) used data from a sequence of courses to explore the relationship between students’ evaluations of teaching and their grades in Introductory Accounting, and their performance in that course as well as their performance in the follow-up course, Intermediate Accounting. They found a positive relationship between students’ evaluations of teaching and their grades in Introductory Accounting, but after controlling for GPA and ACT scores, they found a negative relationship between students’ evaluations of teaching in Introductory Accounting and students’ later performance in Intermediate Accounting. That is, instructors who received higher ratings of effectiveness in Introductory Accounting produced students who actually received lower grades in Intermediate Accounting. These outcomes have since been replicated in other domains (Calculus, Economics, Management, and Law & Management) using stronger designs in which students took standardized tests and were randomly assigned to instructors (Braga et al., 2014; Carrell & West, 2010).

**Biases in Student Evaluations of Teaching**

If student evaluations of teaching do not positively predict learning outcomes, what do these evaluations actually measure? Student evaluations are fairly consistent, in that different students’ evaluations of a given instructor are positively correlated both within the same course (Feldman, 1977; Marsh, 1987; Marsh & Overall, 1979) and over time (Drucker & Remmers, 1951; Howard, Conway, & Maxwell, 1985; Marsh, 1977; Marsh, 2007a; for a meta-analysis, see Feldman, 1989). This consistency suggests that students seem to base their evaluations of instructors on factors that are relatively stable, even if those factors are not strong predictors of learning.

What might those factors be? There is increasing evidence that student evaluations of teaching may be influenced by biases that are unrelated to learning. In an earlier review of this topic—titled, How to improve your teaching evaluations without improving your teaching—Neath (1996) provided 20 “tips” for instructors to boost evaluations without altering their approach to teaching, the top three of which were (1) be male, (2) be organized, and (3) grade leniently.
Subsequent research has revealed that these and other biases are still present in students’ evaluations of teaching. Gender bias is one of the best-documented (for reviews, see Andersen & Miller, 1997; Basow & Martin, 2012). An abundance of correlational data has shown that women typically receive lower evaluations of teaching effectiveness than do men (e.g., Basow, 1995; Basow & Silberg, 1987; Boring, 2017; Boring et al., 2016; Centra & Gauhatz, 2000; Kaschak, 1978; Langbein, 1994; Mengel, Sauermann, & Zolitz, 2017; Mitchell & Martin, 2018; Rivera & Tilsic, 2019; Sidanius & Crane, 1989; Sinclair & Kunda, 2000; Sprague & Massoni, 2005). Even so, this bias is not always present (e.g., Brockx et al., 2011; Feldman, 1992, 1993; Fernandez & Mateo, 1997), and there is some evidence showing that women receive higher evaluations relative to men (Smith, Yoo, Farr, Salmon, & Miller, 2007; Tatro, 1995). Moreover, some research has demonstrated that gender differences in student evaluations may depend on the type of questions being asked. For example, women typically receive higher ratings relative to men on questions related to student-faculty interactions (e.g., Bachen, McLoughlin, & Garcia, 1999; Basow & Montgomery, 2005; Bennett, 1982).

More compelling evidence for gender bias in students’ evaluations comes from creatively designed experiments (e.g., Arbuckle & Williams, 2003; Graves, Hoshino-Browne, & Lio, 2017; Kierstead, D’Agostino, & Dill, 1998; MacNell, Driscoll, & Hunt, 2015). Using an authentic educational context, MacNell et al. asked male and female instructors to teach an online course. Critically, they manipulated the instructor’s perceived gender by leading half of the male instructor’s students to believe he was female, and half of the female instructor’s students to believe she was male. Thus, they used a fully crossed 2 (actual gender) by 2 (perceived gender) design. The instructors only communicated with students via e-mail and through message boards, and they followed the same grading rubric and schedule. After the course, students completed a standard instructor evaluation form. There were no differences in student evaluations of teaching effectiveness according to the actual gender of the instructor. However, there were significant differences according to the perceived gender of the instructor. Students gave higher evaluations to instructors they believed were male relative to instructors they believed were female, regardless of the instructor’s actual gender.

Grading leniency also coincides positively with student evaluations. Instructors who grade leniently typically receive higher ratings than do instructors who grade more conservatively (e.g., DuCette & Kenney, 1982; Eiszler, 2002; Ewing, 2012; Greenwald & Gillmore, 1997a, 1997b; Holmes, 1972; Isely & Singh, 2005; McPherson, 2006; Oliavres, 2001; Stroebel, 2016; but see Heckert, Latier, Ringwald-Burton, & Drazen, 2006; Marsh & Roche, 2000; Palmer et al., 1978). Consistent with this bias, Butcher, McEwan, and Weerahana (2014) reported a naturalistic study in which reductions in grading leniency coincided with decreased instructor ratings. Specifically, after implementing a strict anti-grade inflation policy (i.e., average grades in 100 and 200 level courses could not be higher than a B+ in departments that traditionally provided high grades, instructors’ mean evaluation scores decreased. Of course, without a proper control group, it would be inappropriate to conclude that this policy caused students to provide lower ratings. Even so, this outcome suggests that students may adjust their ratings based on the grades they receive.

Several additional biases have been identified in student evaluations of teaching. For example, an instructor’s age can negatively coincide with the evaluations they receive (e.g., Goebel & Cashen, 1979; Levin, 1988; Peterson, 1980; Sprinkle, 2008). Using a well-controlled experiment, Arbuckle and Williams (2003) had students watch a 30-minute lecture in which slides were accompanied by a stick figure with a gender- and age-neutral voice. After watching the lecture, students were told that the instructor was (a) a young female, (b) a young male, (c) an older female, or (d) an older male. Finally, students took a test over the lecture content. Test scores did not differ between the groups. Even so, on evaluation items relating to the instructor’s enthusiasm and the use of a meaningful tone of voice, students demonstrated a significant gender effect (i.e., students who believed the instructor was male provided higher ratings than did students who believed the instructor was female) and an age effect (i.e., students who believed the instructor was younger provided higher ratings than did students who believed the instructor was older).

Additional biases have been documented with respect to an instructor’s background and appearance. For instructors in the United States, non-Caucasian instructors tend to receive lower evaluations than do Caucasian instructors (e.g., Bavis, Madera, & Hebl, 2010; Ho, Thomsen, & Sidanius, 2009; Littleford, Ong, Tseng, Millikken, & Humy, 2010; Reid, 2010; Smith, 2007). Along similar lines, at least one study has shown that a male instructor who speaks with a non-English accent, compared to a different male instructor presenting the same content with no accent, received lower ratings of instructional quality by students even though students’ actual learning of the content was not affected by the instructor’s accent (Sanchez & Khan, 2016). In addition, instructors who are perceived as attractive often receive higher evaluations than do instructors who are perceived as unattractive (e.g., Felton, Koper, Mitchell, & Stinson, 2008; Felton, Mitchell, & Stinson, 2004; Goebel & Cashen, 1979; Gurung & Vesia, 2007; Hamermesh & Parker, 2005; Rinolo, Johnson, Sherman, & Misso, 2006). As a final example of biases in teaching evaluations—and one that is more easily manipulated than those previously discussed—instructors can boost their evaluations by providing students with chocolate (Younmans & Jee, 2007). For three courses (two statistics courses and a research methods course), students signed up for one of two discussion sections, and halfway through the course evaluated their instructors during the discussion section. For each course, one discussion section received candy before they filled out their evaluations, and one did not. Whereas students’ grades did not differ between these sections, students who received candy before the evaluation rated the instructor as more effective than did students who did not receive candy.

In summary, student evaluations of teaching can be predicted by factors unrelated to teaching and learning. The internal consistency of these evaluations suggests that these factors appear
to be relatively stable, and more detailed analyses of biases in teaching evaluations suggest that they appear to be related to instructor characteristics that are perceived as generally pleasing or desirable, such as enthusiasm, attractiveness, and grading leniency. To the extent that such factors form the bases for students’ evaluations of teaching effectiveness, questions arise about the use of these measures as valid assessments of teaching quality.

Use of Student Evaluations of Teaching in Education

College and university instructors have a great deal of freedom in how they design and teach their courses. Given the important mission of enhancing students’ academic achievement, regular feedback is necessary to determine whether teaching approaches are meeting this goal. As the direct recipients of the instruction, it is entirely reasonable to expect that students would be the ones in the best position to offer this feedback. As we have seen, however, empirical research has provided a wealth of results showing that students are poor evaluators of their own learning, and that their subjective impressions of teaching effectiveness are vulnerable to many biases that are unrelated to teaching and learning.

These biases need not be intentional, or even conscious. Indeed, surveys of students’ perceptions of teaching evaluations tend to be positive, with students reporting that they consider these evaluations to be important and they take them seriously (Ahmadi, Helms, & Raiszadeh, 2001), that they complete the evaluations honestly and accurately (Dwinell & Higbee, 1993; Kite, Subedi, & Bryant-Lees, 2015), and that they believe their evaluations of instructors are not biased by outside factors such as instructor gender, personality, or grading leniency (Ahmadi et al., 2001; Heine & Maddox, 2009; Kite et al., 2015).

Even so, well-intended evaluators may miss important information. There is evidence that students submit evaluations of teaching even when there is no subjective—or objective—basis for them. For example, Reynolds (1977) found that introductory psychology students provided evaluations of the interest and learning value of 10 guest lectures throughout the course—including evaluations for one lecture that never occurred because it had been canceled. Despite the fact that the lecture never happened, 80% of students evaluated it, with about half of the students indicating that the lecture was average and one quarter of the students indicating that it was good or excellent. More recently, Uijtdehaage and O’Neal (2015) inserted the name of a fictitious professor on the end-of-term teaching evaluations for medical students, who typically had multiple professors teach their classes. Despite the fact that students could choose “Not applicable” for any professors who did not teach them, only 34% of students correctly chose this option, while 66% of students provided an evaluation for the non-existent professor.

Notwithstanding honest oversights or cases of mistaken identity, other evidence suggests that student evaluations of teaching can sometimes contain intentional misinformation. In one survey (Brown, 2008), students reported that the ratings of instructor effectiveness tend to be based on the grades that students are receiving in the course, and that students use these ratings to “get back” at their instructors. In another survey (Clayson & Haley, 2011), 36.5% of students reported personally knowing other students who had submitted false information in teaching evaluations because they disliked an instructor. The same sample of students estimated that over 30% of all teaching evaluations contain false information that is knowingly submitted by students.

Student evaluations are nonetheless commonly used to evaluate the quality of faculty teaching. By the mid-1990s, nearly 90% of colleges and universities throughout the U.S. had incorporated student evaluations of teaching, which have become the most widely used measure of teaching effectiveness (Seldin, 1998; Shao, Anderson, & Newsome, 2007). These evaluations are heavily relied upon for evaluating faculty performance, with some survey studies showing that over 94% of deans and administrators report always using student evaluations as a basis for evaluating faculty teaching (Miller & Seldin, 2014), and more than 80% of administrators report using student evaluations as a basis for decisions about tenure and promotion (Beran, Violato, Kline, & Prideres, 2005). To the extent that student evaluations are the primary source of determining teaching effectiveness, faculty who wish to be hired, retained, and promoted may be pressured to adopt approaches that readily boost student evaluations even if they do not boost student learning.

Do Student Evaluations Incentivize Poor Teaching Practices?

Based on the evidence we have reviewed, some ways to boost students’ evaluations would be to provide enthusiastic and entertaining lectures, limit students’ active involvement in the lessons, and provide generous high marks on graded work. If such approaches inflate subjective impressions of learning without consistently enhancing learning, however, are we doing a disservice to students by using these approaches? Teaching practices that produce illusions of learning would seem to create negative downstream effects by reinforcing the divide between students’ perceptions of learning and their actual learning, and would ultimately perpetuate or worsen the poor metacognitive skills that already leave students confused over exam scores and questioning whether they should stay in college.

Abundant research from the learning sciences has provided evidence of concrete approaches to teaching that can effectively enhance student learning and metacognitive skills. These approaches, however, are not likely to boost student evaluations of teaching effectiveness. In particular, one thing we know about effective learning techniques is that they do not always feel effective to students. Techniques that have been shown to significantly boost student learning in academic settings—such as retrieval practice (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013), distributed or “spaced” practice (Carpenter, 2017), and active learning (Deslauriers et al., 2019)—are highly effective yet not always recognized as such. Students often strongly express that they do not learn from these techniques, and believe they learn better from more passive approaches that require less effort (Bjork, Dunlosky, & Kornell, 2013). Indeed, recent
research has identified the perceived effort invested in a learning task as a key contributor to students’ misjudgments of their own learning, in that the effort required is often misinterpreted by students as a sign that they are failing to learn (Kirk-Johnson, Galla, & Fraundorf, 2019).

That is not to say that student’s subjective impressions of learning always run counter to their actual learning. There are poor teaching practices (such as failing to provide clear feedback) that are likely to decrease both perceived learning and actual learning. Though some types of effortful engagement enhance learning (i.e., “desirable difficulties”; Bjork et al., 2013), the investment of effort in general does not automatically lead to better learning. Some types of difficulties are desirable, whereas others (such as struggling to understand one’s own mistakes in the absence of feedback or studying in the presence of distractions) may not promote learning at all and could even undermine it.

To the extent that effective learning techniques are perceived by students as ineffective, instructors who incorporate these techniques would be expected to produce more successful students but might themselves be deemed less effective instructors. The current reliance on student evaluations as a primary measure of teaching effectiveness could incentivize teaching practices that enhance subjective impressions of learning but that are unrelated to actual learning, whereas practices that promote successful learning, but that decrease subjective impressions of learning, are disincentivized. Does this make it risky for instructors to use effective learning techniques? Particularly early in their careers and in teaching-focused positions, instructors may find themselves faced with the difficult decision of whether to incorporate teaching practices that gain them recognition as effective instructors, even if such practices do not positively impact students’ learning.

Can we have the best of both worlds? Is it possible to encourage effective teaching practices that enhance student learning while at the same time accurately evaluate the quality of these practices in a way that rewards good instructors? We turn next to a discussion about potential ways of improving measures of teaching effectiveness.

Improvements and Alternatives to Student Evaluations of Teaching

As we have seen, student evaluations of teaching can be unreliable indicators of learning and are influenced by a number of biases. Because these biases stem from reliance on factors that are not diagnostic of actual learning, a general route to improving the validity of these measures might be to try to decrease the influence of these extraneous factors on measurements of teaching effectiveness. We next consider a variety of approaches for improving upon and supplementing student evaluations of teaching. Though no singular approach is ideal and each has its own limitations, these options allow multiple measures that could provide a more comprehensive approach to evaluating the quality of teaching.

One approach is to carefully consider the construction and implementation of the evaluation questions themselves (e.g., Emery, Kramer, & Tian, 2003; Wright, 2006). From a psychometric perspective, improving upon individual items of a measurement instrument will increase the validity of it. Are students tasked with making assessments of their instructor or learning experience that are beyond the scope of their capabilities? For instance, most students do not have the requisite knowledge about an instructor or the field of study to accurately evaluate his or her knowledge in that field. Eliminating such items may improve the validity of the evaluation instruments. Even so, even if every item is scrutinized, students could still hold misguided beliefs about learning that may (even unintentionally) mislead their responses.

Students’ qualitative comments might be valued over and above their numeric ratings of teaching effectiveness. In doing so, students’ experiences would be considered while minimizing the reliance on a numeric rating that can be biased. However, to interpret students’ responses effectively, those utilizing this information would need training in how to appropriately treat qualitative data, and this may be time-consuming and labor-intensive. As well, little is known about the degree to which students’ responses to open-ended questions are biased by similar factors as are their numeric ratings. In the phantom instructor studies (Reynolds, 1977; Uijtdehaage & O’Neal, 2015), students provided comments for instructors who never existed, suggesting that qualitative responses can be subject to error and bias just like numeric ratings.

Faulty memory can be a source of bias in any type of evaluation. Standard practice is for students to complete evaluations of instructors at the end of the academic term, which means that they must remember their experiences in the course over multiple weeks or months. Regardless of the effectiveness of the instructor, memory itself is subject to a number of biases and heuristics, including high and low points, misinformation, and opinions of others (Schacter, 2008). Thus, one reason why student evaluations are sensitive to external biasing factors such as gender, grading leniency, and even chocolate, could be because these factors are readily available and do not require relying on memory of detailed experiences about the course and instructor, which is cognitively demanding.

Completing evaluations at multiple times during the academic term provides an alternative that is less subject to memory biases and provides multiple measures to be evaluated for reliability. Students’ evaluations collected on individual days and explored for consistency across the term may also reduce concerns about the impact of any one class event (e.g., doing poorly on an exam, getting chocolate from the instructor) on the outcome of those evaluations. One way to carry out this method is to have students rate instructors (i.e., in the standard way) and instructors rate themselves at multiple points throughout the academic term. For instance, Drews, Burroughs, and Nokovich (1987) had students and instructors complete evaluations on 15 pseudo-randomly selected days. Students’ and instructors’ ratings were positively correlated and weak-to-moderate in magnitude (mean $r = .28$). In subsequent research, however, the relationship between students’ and instructors’ ratings depended on student characteristics, in that ratings from advanced students were more strongly correlated with
instructors’ self-ratings than were those from less advanced students, and ratings from students who identified as Caucasian (for instructors also identifying as Caucasian) were more strongly correlated with instructors’ self-ratings relative to those of minority students (Cain, Wilkowskis, Barlett, Boyle, & Meier, 2018). Thus, obtaining multiple evaluations throughout the term may not provide a good measure of teaching effectiveness for all students, and may also be subject to similar biases (e.g., instructor gender, age) as found in the standard approach in which they are only administered once at the end of the term. As well, even though instructors’ self-evaluations provide a standard by which to compare students’ evaluations, instructors’ evaluations are also subjective and may inaccurately reflect students’ learning experiences.

Considering how student evaluations of teaching are constructed, which data are emphasized, and how they are completed by students may lead to measurement improvement. Even with these improvements, however, these evaluations may continue to be poor measures of teaching effectiveness. Thus, it has been argued that student evaluations should never be the sole measure of teaching effectiveness, and alternatives to these measures should be used as well (Emery et al., 2003; Koon & Murray, 1995; Wright, 2006).

One alternative is for peers to observe an instructor’s teaching and provide a written evaluation (Baldwin & Blattner, 2003; for a review, see Koon & Murray, 1995). The peers are other instructors who are experienced at teaching and knowledgeable about the curricular goals of the department and institution, thus ensuring that the instructor is evaluated with the relevant criteria in mind. These evaluations are not free from biases and external influences, however, in part because the instructors being observed might (knowingly or unknowingly) alter their behaviors as a function of being observed. The qualitative nature of these observations might also introduce inconsistencies that are difficult to interpret. For example, Centra (1975) asked peers to observe instructors so that each instructor was observed twice by three peers and also evaluated by students. Whereas multiple observations by the same peer were highly correlated (i.e., observation 1 and observation 2 from peer A), observations from different peers (i.e., observation 1 from peer A and observation 1 from peer B) were weak-to-moderately correlated. As well, peers’ ratings showed little convergence with students’ ratings. Indeed, other researchers have found no meaningful correlation between peer observations of teaching and other measures of teaching effectiveness (Centra, 1975; Howard et al., 1985; Morsh, Burgess, & Smith, 1956; for a review see Marsh, 1987).

Student interviews might provide another means of evaluating the quality of instruction (e.g., Emery et al., 2003; Wright, 2006). If students can provide confidential (rather than anonymous) comments to be reviewed by a faculty member or department chair who is not the instructor of their course, this may improve accountability in reporting and provide the opportunity to follow up on those comments. Although selective removal of anonymity might improve efforts to provide honest comments and reduce intentional misreporting (e.g., Clayson & Haley, 2011), such qualitative responses would likely still be subject to the same measurement issues described above.

Information about the quality of an instructor’s teaching might also be made available in a teaching portfolio (Baldwin & Blattner, 2003). A collection of information and examples that represent one’s approach to teaching—including one’s teaching philosophy, syllabi, example lessons, assignments, and grading rubrics—can provide a comprehensive picture of the teaching practices used by instructors. Such portfolios can be reviewed and evaluated for clarity, fairness, the use of evidence-based practices, and the degree to which an instructor’s approach aligns with the instructional goals of a department and institution.

Research from the learning sciences might inform additional approaches that could supplement the information gained from student evaluations of teaching. In order to measure the knowledge and skills that students have gained from a given course and instructor, it may be useful to conduct follow-up assessments after the course has ended. After completing all of their exams and assignments for the course, students could be contacted at a later time to answer questions probing their long-term retention of what they had learned. The standard practice of evaluating students’ learning at the end of a course may not provide the most comprehensive measure of learning, given that the factors responsible for enhancing long-term durability of knowledge do not always manifest in visible short-term benefits (Soderstrom & Bjork, 2015).

Though follow-up assessments have the potential to capture a more complete picture of students’ durable learning, there are limitations to this approach as well. Performance on these assessments would presumably be sensitive to the nature of the tests and by whom they are created and evaluated, introducing potential subjectivity biases as with course grades. Though standardized measures of desired learning outcomes could be developed, instructors may still feel pressured to “teach to the test” and place greater emphasis on the content that they know will be included on the follow-up assessment. Finally, difficulties could arise in interpreting the outcomes of long-term assessments, which may show a higher-than-desired amount of forgetting of students’ course knowledge. Forgetting has been well-documented for over 100 years (Ebbinghaus, 1885/1913) and commonly occurs for students’ knowledge of course information (e.g., Bahrick, 1984; Conway, Cohen, & Stanhope, 1992). However, when used as a means of evaluating teaching, the visible decline in students’ knowledge over time could raise concerns, especially if comparing students’ performance at the end of a course (potentially based on inflated grades) with their much lower performance on follow-up assessments weeks or months later. Such follow-up measures would thus require careful consideration in their planning and construction, and in the interpretation of their outcomes.

As with any craft, teaching is a multi-faceted skill that requires a multifaceted approach to evaluation. No single measure is capable of reflecting a complete picture of the effectiveness of a given instructor’s teaching. Just as research activity is commonly evaluated via multiple means (for example, both the number and quality of peer-reviewed publications, number of citations, external grants, and impact on real-world policies and procedures), the quality of teaching also reflects a complex
combination of factors that are likely best measured through multiple approaches.

Of course, the utility of any approach depends on what it is designed to measure. In considering whether a given instructor is effective, we have to ask, “effective for what?” This question compels us to define the observable outcomes of effective teaching, at which point it becomes apparent that there is a great deal of variation in the degree to which each of these outcomes is directly measurable. While certain things are readily observable—such as whether an instructor arrives to class on time, shows up for office hours, and responds to students’ questions—others are more difficult to measure. In particular, complexities arise in the measurement of student learning. The counterintuitive and non-immediate nature of learning, along with its sensitivity to grade inflation and other biases, render it a complex construct that may be particularly vulnerable to oversimplification with our current evaluation methods. Thus, whereas some outcomes (such as an instructor’s timeliness) may be straightforwardly captured through a questionnaire response, learning and other complex outcomes are likely best represented through the collection of multiple measures (grades, follow-up assessments, performance in later courses) that provide a more comprehensive representation of the construct itself, and in turn a more accurate picture of its role as an outcome of effective teaching.

Conclusions and Future Challenges

Controversy over the validity of student evaluations of teaching has been the topic of many discussions that will likely continue as long as these instruments are in wide use as a primary tool to evaluate the quality of instruction. It has long been suggested—and empirical research continues to confirm—that these subjective measures are sensitive to error and a number of biases that, even if instructors were inclined and highly motivated, can be difficult or impossible to “improve” upon (e.g., gender, age, attractiveness).

We propose that faulty metacognition is a key contributor to the problem. Students’ misevaluations of teaching effectiveness can be driven by the same factors that underlie their misjudgments of their own learning. Although they do not enhance student learning and can even impair it, teaching approaches that minimize effort and create the appearance of a smooth, well-polished, fluent, and enthusiastic instructor readily boost students’ subjective impressions of what they have learned and their perceived effectiveness of that instructor. Because these subjective impressions are the primary basis for determining teaching effectiveness, and as such are a key metric used for decisions about hiring and promotion, instructors are currently incentivized to adopt teaching approaches that may produce illusions of learning that boost their ratings but can actually undermine students’ learning.

The use of these approaches contributes to the faulty metacognition that produced the biased ratings in the first place. Beyond the problem of failing to equip students with effective skills on how to learn and successfully manage their own learning (which some might consider to be important goals of a college education), perpetuating these illusions of learning introduces future uncertainties as to how such a system can sustain itself. The ever-rising inflation of grades could reflect the increasing expectations of faculty to document evidence of “effective teaching” based on student evaluations (e.g., Stroebe, 2016). Such a trend can only continue for so long, however, before all students reach the top of the grading scale. As well, there are upper limits on the amount of enthusiasm, attractiveness, and chocolate that instructors can provide.

Thus, a future challenge that we foresee in education is an inevitable re-thinking of how teaching effectiveness is conceptualized and measured. It is hard to predict when this paradigm shift will occur, but chances are that it will be brought on by the decreasing functionality of a current system that will one day cease to provide interpretable data because of the numerous factors that can artificially inflate measures of effective teaching. It is hoped that future conceptualizations of teaching effectiveness include research-based evidence for improving student learning and metacognition as a strong basis in formulating measurements that accurately and reliably reflect the quality of teaching.

Author Contributions

All three authors contributed to the idea, conducted literature reviews, wrote sections of the manuscript, and provided edits to the manuscript prior to submission.

Conflict of Interest

The authors declare that they have no conflict of interest.

References


Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students’ evaluations of their instructor. *Journal of Educational Psychology, 63*, 130–133.


Hot or not: Do professors perceived as physically attractive receive higher student evaluations? The Journal of General Psychology, 137, 19–35.


Received 11 November 2019; received in revised form 29 December 2019; accepted 29 December 2019. Available online 12 February 2020.