Challenging Misconceptions About Student Ratings of Instruction

IDEA Paper #58 • April 2016

Stephen L. Benton and Kenneth R. Ryalls • The IDEA Center

Abstract

Data from student ratings of instruction (SRI) are used ubiquitously as a key element in providing instructors with valuable feedback and evaluators with critical student input. Nonetheless, calls for the elimination of SRI continue to appear in academic journals and higher education periodicals. This paper brings to bear the huge body of research on SRI to which so many academics and institutions have contributed. Some of the most egregiously erroneous statements about SRI are rebutted with brief reviews of the readily available compelling evidence. Although some faculty frustrations about misuse of SRI are valid, we argue that inclusion of student voice is critical. Students can provide useful feedback because they have firsthand experience over multiple occasions of what actually occurred in the classroom. Recommendations are made for best practices in using SRI as one of many sources for improving and evaluating teaching.

The fact that an opinion has been widely held is no evidence whatever that it is not utterly absurd; indeed, . . . a widely spread belief is more likely to be foolish than sensible.

— Bertrand Russell, Marriage and Morals

Institutions of higher education believe that improving teacher quality and effectiveness are priorities. These institutions know that perception of the quality of teaching is a principal reason students give for selecting a college or university (Shah, Nair, & Bennett, 2013), and institutions spend thousands of hours of faculty time and considerable financial resources to measure and improve student achievement (Cooper & Terrell, 2013). When making tenure decisions departments recognize faculty research in pedagogy and evaluation as legitimate academic discourse, as reflected in the many books focused on teaching and learning that appear every year. For example, “in the past 10 years (2015 to 2005) Wiley published 646 books in its Higher and Adult Education Division; . . . of those, 175 were specifically in the Teaching, Learning, and Curriculum category” (C. Allard, personal communication, January 8, 2016). Academics, departments, and colleges accept, then, that teacher evaluation is integral to their missions.

Data from student ratings of instruction (SRI) (aka “student evaluations of teaching” or “course evaluations”) are used ubiquitously as a key element in measuring teacher effectiveness. Indeed, student voice has been argued as essential to positive change in the classroom, because students can provide critical information on the improvement of teaching and learning (Quaglia & Corso, 2014). SRI are often part of decisions about merit salary, tenure, promotion, and helping faculty improve courses and instruction. Because of their widespread use—and, at times, misuse—SRI have understandably undergone extensive scrutiny with upwards of over 3,000 publications devoted to them (Benton & Cashin, 2014). The topic has been studied extensively, perhaps more than any other in higher education. Why, then, do some institutions get teaching evaluation so wrong and so many misconceptions and misunderstandings persist about the validity and reliability of SRI?

Claims of bias (non-instructional factors influencing evaluations) continue to appear in academic journals (e.g., Boring, Ottoboni, & Stark, 2016; MacNell, Driscoll, & Hunt, 2014; Stark & Freishtat, 2014) and higher education periodicals (e.g., Asher, 2013; Berrett, 2015; Flaherty, 2016; Mulhere, 2014; Wieman, 2015; Zimmerman, 2014). Such claims frustrate and disappoint those familiar with the vast research literature providing empirical evidence that supports SRI utility (validity) and consistency (reliability) (see Benton & Cashin, 2011, 2014; Hativa, 2013b; Marsh, 2007; Theall & Franklin, 2001, for reviews). Consequently, one purpose here is to address commonly held misbeliefs about SRI by bringing to bear the huge body of research and data to which so many academics and institutions have contributed. We begin by citing papers containing some of the most egregiously erroneous statements about SRI and rebutting those statements with brief reviews of the readily available compelling evidence. Next, given how frequently administrators and promotion/tenure committees misuse SRI and the deleterious consequences of doing so, we acknowledge and address faculty complaints about SRI. Then, we recommend ways SRI can be used to inform teachers and identify areas needing improvement within an institution. We end with a brief review of what the research indicates are current best practices in using SRI effectively in evaluating teaching.

Flagrantly False Claims About SRI

Bad Teachers Get Better Evaluations

Among the more egregious claims is that "professors who receive high evaluations are worse teachers than their peers" (Zimmerman, 2014). In a similar vein, Stark and Freishtat (2014) voiced, "Good teachers can get bad evaluations." Both of these statements miss the point—SRI are designed to measure teaching effectiveness *in a given course*, not teaching effectiveness in general. Certainly a usually effective instructor can get a bad evaluation in a given course. The real question should be how do the ratings look across *multiple* courses? Research has established that at least six to eight class ratings should be collected before reliable, summative decisions about an individual's teaching effectiveness can be made (Benton, Li, Brown, Guo, & Sullivan, 2015). So, yes, a good teacher can get bad evaluations on occasion, often due to personal or environmental factors influencing performance. But how does that mean that SRI are unreliable and should not be trusted? A good teacher who gets one poor evaluation across eight different classes is obviously a good teacher. Likewise, a teacher who gets poor evaluations across eight classes with only one good evaluation is, logically, in need of help and improvement. The odd poor evaluation for a good teacher is analogous to the coldest day argument by climate change deniers; an outlying single experience should not be compelling evidence.

If the incidents of good teachers getting bad evaluations were widespread, we would not expect to find the consistent positive relationships between SRI and other indicators of teaching effectiveness. In point of fact, SRI correlate positively with:

1. students' actual achievement in the course as measured by exam performance (Beleche, Fairris, & Marks, 2012; Benton, Duchon, & Pallett, 2011),
2. instructor self-ratings (Feldman, 1989a; Marsh, Overall, & Kesler, 1979; Marsh & Dunkin, 1997), and
3. ratings by colleagues, administrators (Feldman, 1989a), and trained observers (Feldman, 1989a; Marsh & Dunkin, 1997; Murray, 1983).

So, the comments claiming good teachers do not consistently receive the highest SRI are simply not true and not supportable based on research. Ratings from teachers themselves, their administrators, colleagues, alumni, and trained observers confirm the validity of SRI.

"Tough" Demanding Teachers Receive Lower SRI

This old saw has been around since the first days of SRI, and, unfortunately, gets repeated yet today along with its corollary that "easier" teachers get higher SRI. Or, as Lyell Asher (2013) asserted in an article published in the *Wall Street Journal*, "easing up on demands and raising grades will get you better reviews [student ratings] at the end" (para. 4). This assumption that students are out for the easy "A" is insulting to the vast numbers of students who are working hard to gain an education and is perhaps indicative of a general attitude toward students from those arguing vehemently against SRI use.

The cynical assumptions underpinning these types of assertions are that students are interested only in grades and do not want to be challenged in their educations. But in a study involving over 50,000 classes across eight academic disciplines, Centra (2003) found that the grade students expected to earn was only weakly related to SRI. Others have similarly reported low, positive correlations (Braskamp & Ory, 1994; Centra, 2003; Feldman, 1976; Howard & Maxwell, 1980, 1982; Marsh & Dunkin, 1997; Marsh & Roche, 2000). Even this low positive correlation between grades and ratings may not necessarily indicate instructors are lowering standards to get higher ratings. It could well indicate that students who learn more earn higher grades and assign higher ratings, which supports the validity of SRI. A third possibility is that student characteristics, such as motivation and interest in the subject matter, could lead to greater learning and, therefore, higher grades and student ratings (McKeachie, 1997).

Evidence shows that Asher's assertion about grade leniency is not only wrongheaded but perhaps actually the inverse of the truth. If teachers really want to improve course ratings, they would do well to practice other more productive behaviors than assigning lenient grades. Challenging students, stimulating their interests (Marsh & Roche, 2000), and making appropriate changes to instruction and the course based on student feedback (Centra, 2003) are more likely than leniency to lead to higher SRI and greater student learning. Moreover, research conducted in nearly 500,000 classes across more than 300 institutions revealed that instructors are more likely to earn high SRI when their students say their teacher challenged them and had high achievement standards (Benton, Guo, Li, & Gross, 2013). Let us emphasize: a half million classes and 300 institutions. It's time for SRI deniers to acknowledge that the research, the science, is simply not on their side.

Students Are Not Qualified to Judge Teaching Effectiveness

Arguing that the worst or most lenient teachers receive the highest ratings rests upon another fraudulent assumption: students are not qualified to judge teaching effectiveness. Understandably, we may be put off when the evaluators of our work are less educated than we are. How can we trust undergraduate students to render valid judgments about our teaching effectiveness when most of them have never taught? Perhaps an analogy from another profession may be of use in answering this question. Typically, hospital administrators are expected to evaluate the effectiveness of physicians, a practice many who have had experience with doctor visits would support, as a patient's voice should matter in decisions about improvement of care. Logically, one important factor in physician evaluation would be patient ratings of experiences with their physician that include perceptions of progress in recovery, the physician's interpersonal skills, quality of care, and so forth (Manary, Boulding, Staelin, & Glickman, 2013). Patients are not doctors, but ignoring the input of patients about their doctors would be foolish. Students are not professors, but the same logic applies—ignoring student input is foolish. If patient input were the only source of evidence in making decisions about which doctors were effective, there would justifiably be concern. On the other hand, if we decided to ignore patient perceptions, we would lose out on some valuable information about how to improve medical services just as we lose valuable information about improving teaching if we ignore student feedback. Yet the old chestnut about students' being incapable of providing useful feedback still finds its way into print even today.

The time and resources devoted to debating whether students are capable could be better spent constructing the questions students are asked to answer. Students know how to assess if the right questions are asked. In fact, in a review of 31 studies Feldman (1989a) found that student views of what constitutes effective teaching are very similar to those of faculty (average correlation = .71). In the realm of patient health-care assessment, where some physicians and academics argue patient feedback is not credible because patients lack formal medical education, Manary and colleagues (Manary et al., 2013, para. 1) get it right: "when designed and administered appropriately, patient-experience surveys provide robust measures of quality." The same can be said for SRI.

SRI Are Unreliable

Reliability refers to consistency, and well-constructed SRI have a great deal of it (see review by Benton & Cashin, 2014). Actually, SRI within the same class tend to be highly consistent in students' own ratings, in ratings over students within the same class, and in ratings of the same instructor across multiple courses.

Perhaps the unreliability assertion is getting passed around again for two reasons. First, many home-grown ratings systems are poorly designed. Just as a camel is a horse designed by a committee, so goes the plight of many SRI surveys faculty committees create (Hativa, 2013a). Another possible explanation is the increase in web-based SRI, which typically have lower response rates than those administered on paper. The thinking may be that ratings based on low student response rates cannot be trusted. Certainly, ratings based on lower response rates cannot be assumed to represent the overall class perceptions as well as higher response rates. But, representativeness is a different issue than reliability. The former is tied to the percent of students in the class that respond—the greater the response rate, the more representative are the scores derived from the course rating. Reliability, on the other hand, is related to *sample size*, or the number of student raters. If 50 students out of a class of 100 responded to a survey, their ratings would be more statistically reliable than if 19 students out of a class of 20 responded even though the 19 responders would be more representative.

So even though the reliability of any measure does increase as the number of observations increase, it does not follow that a low number of observations means those observations are not reliable; even classes with low response rates can provide useful information for a teacher. Perhaps an analogy is helpful. If one person tells us we have a tail, we can write it off as ridiculous. If a second person says, "Hey, you have a tail," they have our attention, but we can still think, "These people are clearly mistaken." But, if a third person on a different occasion pulls us aside and says, "Hey, did you know you have a tail?" we still might not necessarily conclude we have a tail, but we should certainly be concerned that folks are telling us that we do. The same goes for ratings, even those with low response rates. The first time we get "bad" ratings from a course with a low response rate, we can perhaps pay little attention to it. Perhaps it is an instance mentioned before of a good teacher getting an occasional poor rating. But, across time if the ratings are consistently low, even from classes in which few actually participate in the rating, we need to respond, gather additional information (e.g., from peers), and see what we can do to improve the situation.

Because well-constructed SRI present multiple information from individuals (students) within a class and are collected across multiple occasions, one can make the case that students provide the most reliable source of feedback about teaching (Marsh, 2007). In contrast, class observations performed by an administrator or a peer—be they trained or untrained evaluators—typically represent only one observation on one occasion. In this case we do not know what the consistency/reliability is of their ratings. For reliability, trust SRI.

Personal Factors Unrelated to Learning Influence Ratings

Another favorite claim put forth yet unsupported by research is that overall ratings are strongly influenced by instructor characteristics unrelated to student learning, such as instructor gender, ethnicity, and personality (Stark & Freishtat, 2014). For example, the popular press gave unjustifiable attention to two studies that claimed gender bias in SRI (Boring et al., 2016; MacNell et al., 2014), which were refuted in an editorial note (Ryalls, Benton, Barr, and Li, 2016). In studies of SRI collected in actual classes, gender is only weakly related to ratings (see literature reviews by Benton & Cashin, 2011, 2014). Further, Li, Benton, and Ryalls (in press) analyzed data collected from IDEA SRI in over 15,000 classes taught by female instructors and over 12,000 taught by male instructors. Multiple institutions, Carnegie classifications, and disciplines were represented. The authors found no differences in overall ratings of teaching, the course, and average student progress on relevant learning objectives.

When differences between male and female instructors have been found, they have typically occurred in research laboratory studies (where students rated descriptions of fictitious teachers who varied in gender). However, in studies of ratings of actual teachers there is only a very weak relationship that favors female instructors (Centra, 2009; Feldman, 1993). In reviewing several studies, Feldman (1992) concluded, "Any predispositions of students in the social laboratory to view male and female college teachers in certain ways (or the lack of such predispositions) may be modified by students' actual experiences with their teachers in the classroom or lecture hall" (Feldman, 1992, p. 152). Feldman's point corresponds to Gordon Allport's (1954) contact theory, which suggests stereotypes can be overridden by actual personal interaction, a hypothesis well-supported according to Pettigrew & Tropp's (2006) meta-analysis of 515 studies. Even electronic contact (both text-based and video-based online interactions) has been shown to reduce prejudice (Amichai-Hamburger & McKenna, 2006).

With respect to race/ethnicity, very few studies of actual ratings of instructors in the classroom have been conducted (for reviews, see Benton & Cashin, 2014; Centra, 1993; Huston, 2005). Some studies have randomly assigned students to rate a fictitious instructor on various qualities based on changes in the instructor's name, which implied a certain gender and ethnicity (Smith & Anderson, 2005; Anderson & Smith, 2005). Others were based on a computer-animated professor who varied in gender and race (Basow, Codos, & Martin, 2013).

Conflicting conclusions were reported in two different studies conducted on local ratings at single institutions. In one, Black instructors received significantly lower scores than White and "Other" faculty on overall ratings of the course and teaching effectiveness (Smith, 2007). In the other study, no support was found for the initial hypothesis that students would rate minority faculty lower than majority faculty (Ludwig & Meacham, 1997). Clearly, given the limited number of studies and the conflicting outcomes, more research is needed. Racial bias cuts across all facets of society, and administrators and faculty are professionally responsible for examining the extent to which it exists at their institution. One of their most useful sources of information can be analyses of SRI in their local setting. When combined with other indicators of teaching effectiveness, SRI can be a great help in decision-making.

The conclusion from studies of instructor personality is that it has little impact on ratings (Braskamp & Ory, 1994; Centra, 1993). The few personality traits related to SRI, such as positive self-esteem, energy/enthusiasm, and orderliness, tend to enhance teaching effectiveness and are, therefore, not considered biases. Displaying energy and enthusiasm could stimulate student interest, which is positively correlated with multiple learning outcomes. Being orderly creates classroom structure, which is associated with greater student learning of cognitive outcomes (Benton et al., 2015). It is, therefore, not so much the personality of the instructor that matters as much as the personal characteristics manifested in the classroom. Most of the relationship between instructor personality and SRI is most likely connected to the behaviors the instructor displays in the classroom. Or as Braskamp and Ory (1994, p. 180) concluded, the influence of personality "may be caused more by what [instructors] do in their teaching than by who they are."

To say that teacher gender, race, and personality do not exert great influence on SRI is not to deny that bias does exist for some students in a class. Of course bias exists to some degree in student feedback, as course ratings are surveys designed and filled out by humans. But, bias in student ratings due to these instructor variables is not large and should not greatly affect teaching evaluations. Moreover, faculty need to understand and be comfortable that SRI are robust against other potential biases. For example, SRI are not strongly related to instructor age and teaching experience (Marsh & Hocevar, 1991). They are also robust against potential biases brought on by some student characteristics. Student gender is not highly correlated with ratings, although student-gender-by-instructor-gender interactions have been reported. In a study involving a large number of two- and four-year institutions across a variety of academic disciplines, Centra and Gaubatz (2000) found female students gave slightly higher ratings to female instructors. However, the authors did not consider the differences large enough to affect personnel decisions. Other student characteristics, including year in school (Davis, 2009; McKeachie, 1997), grade-point average (Braskamp & Ory, 1994; Centra, 1993; Marsh & Dunkin, 1997; Marsh & Roche, 2000; McKeachie, 1997), and personality (Abrami, Perry, & Leventhal, 1982) have little or no relationship to SRI. Likewise, SRI are not meaningfully affected by the time of day the course is offered (Aleamoni, 1981; Feldman, 1979) and by whether they are administered online versus on paper (Benton, Webster, Gross, and Pallett, 2010a) or in online versus face-to-face courses (Benton, Webster, Gross, and Pallett, 2010b).

To point out bias in a rating given by a human and use it to negate the usefulness of that rating makes no sense. If the efficacy of student feedback is to be ignored because of bias, then one must also throw out peer and administrator feedback, as well as Promotion and Tenure Committees, annual reviews, reference letters, instructor self-reflections, and anything else that has a human element to it. Grades given by instructors would also be useless, as instructors are human and therefore full of bias. The question then is not "Is there bias in this tool?" but "Can we find usefulness in these data in spite of the bias inherent in humans?" The answer to this question is yes, provided the survey instrument is well designed.

Students Tend to be Motivated More by Anger About a Low Grade than Satisfaction

This claim assumes, when ratings are not mandatory, that students who are upset about their expected grade are more likely to complete the ratings form than students who are doing well in the course. However, the relationships between response rates and overall ratings of the teacher and course tend to be quite low in surveys administered online and on paper (Benton et al., 2010b) and in classes taught online and face-to-face (Benton et al., 2010a). The assumption that anger or revenge will cause more students to respond than other motivations seems unfounded. High achieving students are more likely than others to respond to an SRI survey (Avery, Bryant, Mathios, Kang, & Bell, 2006; Porter & Umbach, 2006; Porter & Whitcomb, 2005). Nonresponse to online SRI surveys is more common among grade D and F students compared to grades of A, B, and C (Adams & Umbach, 2012). Why would high-achieving students be more likely to respond? The answer may be that high-achieving students have more positive feelings about the course or the institution. In fact, student satisfaction with college is positively correlated with grade-point average (Kuh & Hu, 2001). High achievers are, therefore, more likely than others to respond to an SRI survey not out of anger but out of satisfaction. Also, it makes intuitive sense that D and F students are less likely to have done the work during the semester and are therefore less likely to do the 'work' of providing feedback at semester's end. If they checked out during the semester, the chances of checking back in to rate the instructor is probably small.

Millennial Students Are More Punishing in Their Ratings

Every generation of teachers has probably uttered something like, "The students of today are not like the ones I taught years ago." So, no surprise, the Millennial generation is getting its turn to be criticized. Members of the Millennial generation, ranging in birth dates from the early 1980s to the early 2000s, are believed to share common characteristics, one of which is a sense of entitlement (Nilson, 2013; Twenge, 2006). Nilson (2013) has argued that Millennial students feel entitled to receive high grades without putting out much effort. Accordingly, some teachers fear that Millennial students will be even more likely than prior generations to assign low ratings to instructors who give lower than expected grades. In point of fact, average overall ratings of the instructor and course have increased steadily since 2002 (Benton et al., 2015), which refutes the notion that Millennials tend to be more "punishing" in their ratings.

Some might argue that the gradual inflation in SRI is evidence that instructors are "dumbing down" the curriculum to get high ratings. If that is the case, the fault lies not in the instruments used to assess teaching effectiveness but the system that over-emphasizes them and the faculty who lower their standards. But lowering standards is misguided because, as mentioned previously, ratings of teaching and the course are higher when students report the instructor had **high achievement standards** and expected students to share responsibility for their own learning (Benton et al., 2013). In addition, ratings of teaching effectiveness are higher when instructors encourage students to think for themselves (Zhao & Gallant, 2012).

But, why has there been a steady increase in average ratings since 2002? If one is to believe, as has been argued, that SRI are valid and reliable, what accounts for the fact that Millennials are rating teachers higher than previous generations? One explanation is that students of today are simply more generous. But then if that were the case a high positive correlation between teacher standards and ratings would seem unlikely. Students are discriminating in their ratings of teachers—they do not merely assign high ratings to everyone. A more logical explanation is the success story of the growing emphasis institutions place on teaching. Faculty development is a growing field—recently having gained its own professional journal (*Journal on Centers for Teaching and Learning*) and its own professional organization, the Professional and Organizational Development (POD) Network in Higher Education. Many centers for teaching and learning (CTL), which provide developmental assistance to faculty seeking to improve teaching, are currently thriving (Flaherty, 2014).

Miami University provides a telling example of the impact a CTL has on a campus. In 2003 the university targeted the top 25 introductory courses with the largest enrollments and implemented a course redesign project based on an inquiry-oriented approach to teaching. Emphasis was placed on active learning, peer collaboration, and critical thinking. In time, this approach was generalized across other courses. Between 2003 and 2011 significant increases occurred in course application of theories or concepts to practical problems, student peer engagement, and the amount of preparation students reported doing outside of class (Nadler, Shore, Taylor, & Bakker, 2012). The trend reported at Miami is not isolated. The positive linear trends across time in overall ratings of teaching in the IDEA database have been accompanied by significant and meaningful increases in teaching methods associated with active learning (Benton et al., 2015). The largest increases have been in collaborative learning; involvement of students in hands-on projects such as research, case studies, or real-life activities; and facilitating interactions between students of diverse backgrounds and viewpoints. So the steady increase in ratings found in the Millennial generation is associated with instructor use of more student-centered approaches to teaching. Indeed, SRI deniers who turn their backs on the overwhelming and incredibly useful information generated by SRI (because they deny their reliability) may well be turning their backs on some of the most positive pedagogical techniques emphasized in the past couple of decades.

Summary

The rebuttals made thus far are in response to the most unreasonable indictments of student ratings. Credible evidence is lacking to support the views that bad teachers get better evaluations, having high standards leads to low evaluations, students are not competent to rate, SRI are unreliable, personal factors unrelated to learning strongly influence ratings, students are more motivated to respond out of anger than satisfaction, and Millennials are more punishing in their ratings than previous generations.

Faculty Frustration with SRI Is Understandable

As shown above faculty frustrations with the evaluation process have led to some faulty claims about SRI themselves, but faculty may have legitimate gripes about the process and have just focused on the wrong sources of real problems in evaluation. SRI are not the problem, but faculty should be rightly upset when SRI are misused and overemphasized in summative decisions about teaching effectiveness. One way ratings are misused is when evaluators make too much of too little (Pallett, 2006). Although SRI tend to be highly consistent among students in the same class and across classes for the same instructor (Benton et al., 2015), as with any survey instrument there is “noise” in the data. We should expect ratings collected on consecutive days to vary like blood pressure, not remain constant like height. Yet, administrators may at times make decisions about salary increases based on small differences in mean ratings. Student ratings are overemphasized when they count too much in decisions about salary, promotion, and tenure. The IDEA Center has long recommended ratings should count no more than 30% to 50% of the overall teaching evaluation (Hoyt & Pallett, 1999), which is only one aspect of the “three-legged stool” (i.e., teaching, scholarship, and service) of a holistic evaluation process. Administrators who make SRI the primary or only measure of teaching effectiveness create mistrust and a breeding ground for claims of bias in student ratings. But such misuse should not lead to indictments of SRI but rather of the process. Effective use requires collecting multiple indicators (e.g., peer ratings, artifacts) of teaching effectiveness.

We demonstrated above that students are qualified to provide useful reliable feedback on teacher effectiveness, but there are many elements of teaching about which students are not qualified to judge: course objectives, the instructor’s subject-matter knowledge, assessments and grading standards (for more information on students’ appropriate role in teaching evaluation, see Arreola, 2006, and Hoyt & Pallett, 1999). Evaluation processes which confuse those aspects of teaching about which students are qualified to respond and those about which student input is not useful can and do add to faculty frustrations. Designers of teaching evaluation processes need to take care in constructing ratings forms, limit forms to those aspects of teaching about which students are capable of sound judgment, and make SRI only a part of a larger evaluation process. Good surveys are difficult to design, and institutions that take on the burden of self-designing their own SRI do so at their own peril. Those institutions who employ an instrument designed by a committee decades ago, or worse yet allow each department to develop its own tool, are at risk of making decisions based on questionable data.

Faculty perhaps look to blame SRI for their unhappiness about evaluation because the evaluation process is not emotionally neutral. Even when ratings are used properly, the process of being evaluated can create anxiety which could negatively affect motivation to teach as well as motivation to try innovative methods (Theall & Franklin, 2001). For example, feedback from SRI could affect *teacher self-efficacy*, the belief in one’s ability to help students succeed. Self-efficacy is important because students achieve more when teachers believe they can make a difference in their students’ lives (Skaalvik & Skaalvik, 2008; Ware & Kitsantas, 2007). When they have high self-efficacy, instructors are more willing to try out new teaching strategies, set high achievement goals for their students, put forth effort, and persist in teaching (Ormrod, 2014). Fear of how teaching behaviors could affect ratings might create reluctance to take risks in the classroom or to cultivate high expectations. Teachers may fear innovation will be “punished” by low ratings. In contrast, teacher self-efficacy flourishes in contexts where feedback draws attention to effective teaching behaviors and provides constructive, specific recommendations for improvement (Hoy, 2000). Hence IDEA’s *Diagnostic Feedback* report identifies teaching methods as either “strengths to retain” or ones to consider increasing. The effect of feedback on self-efficacy is enhanced when instructors engage in discussion with an administrator or peer and feel free to ask for help (Hoy, 2000; Finnegan, 2013). Moreover, innovation is more likely when instructors believe their attempts to engage students in new methods will be rewarded, not punished.

Many times faculty blame the messenger (SRI) for factors faculty feel are beyond their control. If no remedies exist to help with low SRI, then one’s efforts might just as well be directed against the rating instrument itself. Faculty sometimes make faulty *attributions*, causal explanations to explain success and failure. When they attribute low ratings to a faulty evaluation system, bias in the ratings, or student incompetence, instructors are likely to become angry at the evaluation system itself (Hareli & Weiner, 2002). And that anger can intensify when faculty feel trapped. At such times

SRI can also either positively or negatively affect *expectancy*, beliefs about the likelihood of responding successfully, given our current ability level and external events that may hinder performance (Wigfield & Eccles, 2000). One external event is sufficiency of resources and support available to assist us. If faculty receive negative feedback about their teaching but lack developmental resources, they may doubt their chances for improvement. Again, evaluation processes which do not include concrete specific ways to help faculty who wish to improve lead to frustrations. By using multiple measures, rewarding innovative approaches to teaching and employing a holistic evaluation process, faculty irritation about ratings can be diminished.

Given the above discussion we should not be surprised, then, that countless hours and untold dollars have been invested in searching for negative evidence about SRI. In situations where they are used improperly, count too much in the evaluation process, and are not supported with helpful resources, we can simply reiterate that such evidence is not negative about SRI itself but the environment in which they are used. The huge preponderance of evidence shows the utility and reliability of SRI. So, as Theall and Franklin (2001) point out, the time and resources devoted to trying to prove that SRI are biased would be better directed toward developing new strategies for teaching and helping students learn.

Can SRI Really Be Used to Improve Teaching?

With the continual cycle of "end-of-course evaluations," one is tempted to ask, "Are all these ratings really helpful?" And the answer is demonstrably, convincingly, "Yes, provided the instrument asks the right questions." Just as importantly, SRI's helpfulness can be increased depending upon what we do with the feedback we receive. Simply reflecting upon our SRI can make a difference, as a meta-analysis of 17 studies revealed; faculty who reflect on ratings administered midterm improve on end-of-course evaluations (Cohen, 1980). The greatest gains in SRI, however, are found among instructors who combine student feedback with consultation involving a peer or faculty development specialist (Brinko, 1990; Burbano, 1987; Cohen, 1980; Hampton & Reiser, 2004; Hativa, 2013a; Knol, 2013; Marincovich, 1999; Marsh, 2007; Marsh & Roche, 1993; Ory & Ryan, 2001; Penny & Coe, 2004). When combined with consultation, feedback from ratings can have strong effects on the instructor's knowledge, focus on teaching, and plans for improvement. For SRI deniers to debunk the SRI themselves to justify, perhaps, their own low ratings is especially troubling given that SRI yield so much useful information that those very deniers could use to boost their own student ratings of teaching skills and of how much students feel they learned (Knol, 2013).

Discussing ratings with a peer or consultant improves their usefulness, especially when the consultation addresses problems students have identified (Marsh & Roche, 1993). Common areas needing improvement are student-teacher interactions, active learning opportunities, teacher expectations and standards, class preparation, and assessments (McGowan & Graham, 2009).

Mulling over feedback from SRI can be humbling and somewhat frightening. We have to take an honest look at what the students are telling us. If a trusted consultant is unavailable, we can at least engage in self-reflection by comparing students' perspectives with our own experiences planning and teaching the course. If the students' views are substantially different from our own, we can use that cognitive dissonance to take action. How might we teach the course differently next time? Are there simple changes we can make to address the issues raised? Any attempt at improvement is far better than simply placing the SRI class report in a folder and ignoring it until we assemble our teaching portfolios, provided administrators endorse the attempt as part of the process of effective teaching, regardless of its effect on the SRI. Without encouragement to innovate and take risks to improve built into an evaluation system, faculty will be unlikely to attempt new methods.

Beyond the course level, SRI have a place in both assuring the quality of teaching and improving it within an institution (Palermo, 2013). Therefore, institutions should use SRI to identify areas of strengths and weaknesses at the unit and institution levels. Data aggregated across departments, colleges, and/or curricula (e.g., general education) can be analyzed to identify most frequent areas needing improvement. Department heads can then be encouraged to allocate resources for faculty development focused on targeted teaching behaviors and learning outcomes. Students can be informed of programmatic and course changes designed to address the feedback provided on the SRI report (Palermo, 2013).

Best Practices in Using SRI for Teaching Evaluation

When used appropriately, SRI have a place in summative evaluation. Whereas formative evaluation is focused on improvement, summative ratings can provide indirect evidence of instructional effectiveness. Ratings should be collected from every course, but not necessarily every semester, so that evaluators can look for global trends in the data, such as steady performance, declines, or improvements, as well as certain courses that may need attention. In using SRI for personnel decisions, the following best practices are recommended.

Use Multiple Measures

Any evaluation of teaching effectiveness should incorporate multiple measures, such as peer ratings of course goals, design, and assessments; direct student outcome measures (e.g., creations, projects, papers); instructor self-reflections; and so forth (Hoyt & Pallett, 1999). When the combined data are in agreement, reliability increases (Cashin, 1996). Faculty need to be selective and strategic when they assemble multiple information sources (Halonen, Dunn, McCarthy, & Baker, 2012). The source should have actual knowledge of the measure being evaluated (Arreola, 2006). Students, for example, have firsthand knowledge of what occurs in the classroom or in online discussions but are unqualified to render judgments about the goals and content of the course, methods and materials used in teaching, and student evaluation practices (Keig & Waggoner, 1994).

Peer review is a credible source for evaluation of course goals and objectives, intellectual content, methods and materials used in teaching, quality and appropriateness of evaluation practices, and evidence of student learning. But, peer review improves in validity when faculty undergo some training. Ratings by colleagues can be unreliable and less valid when done by untrained observers and with an unsystematic approach to the evaluation (Marsh, 2007; Marsh & Dunkin, 1997). When done in an environment of distrust due to a faulty evaluation process, peer review can become downright protective, with no true feedback that could ever be construed by an administrator as negative. In this worst-case scenario, the peer review is not geared toward teaching development, but is instead written to protect a fellow faculty member from attack. In these cases, the process of evaluation itself needs an overhaul so that peers again feel comfortable in providing meaningful critique, allowing colleagues to assist each other in improving their teaching. *External recognition* from outside experts can provide evidence that the faculty member's teaching is exemplary. Nominations for a teaching award, invitations to write a chapter or book about teaching, and requests to speak about teaching practices or share course materials are examples that the instructor's teaching is praiseworthy. *Embedded assessments* (i.e., student completion of class assignments and activities aligned with learning outcomes) signal accomplishment of specific learning outcomes. Examples include student writing samples, self-reflections on service learning projects, comparisons between students' subject-matter knowledge before and after instruction, and student portfolios of completed work. Participation in *professional development* activities demonstrates the desire to improve when evidenced by instructor self-reflection about how the activity led to modifications in the course or approaches to teaching. *Exemplary contributions to the department* are shown when teaching large sections, taking an extra teaching assignment to accommodate increased enrollments or a sick colleague, developing curriculum and aligning it with accreditation standards, and helping colleagues improve their teaching.

Recognize the Limitations of Each Measure

Each source of evidence has its shortcomings. The limitations of SRI have already been mentioned. We must also be cautious when interpreting information provided by the instructor. Some colleagues may not be comfortable reflecting on their own behavior, and in some cultures instructors might find it inappropriate to speak highly of themselves. Others are exceptionally skilled at self-promotion and at organizing a teaching portfolio (Zakrajsek, 2006), which can make fair assessment sometimes difficult.

With respect to peer review of course materials and embedded assessments, additional cautions must be raised. Faculty may differ in their philosophies and approaches to teaching. A faculty member who takes a teacher-centered approach to instruction might not be the best person to evaluate a colleague who employs a student-centered approach with active learning (Zakrajsek, 2006). For this reason, at least three raters should review any one person's materials. Another shortcoming is that peer review also requires a great commitment of time for training and evaluation. Rewarding faculty for agreeing to evaluate their colleagues is important; recognizing and rewarding peer evaluation as important service to the department and college will encourage this extremely useful activity. Finally, "friends as peer reviewers" should be avoided, because ratings that are glowing across all faculty are next to useless.

External recognition and professional development also have their shortcomings. Evaluators must scrutinize the significance of such achievements and learn as much as they can about awards, nominations, invitations, and developmental opportunities. Was the focus of the professional development aligned with the strategic plan of the institution or department? Was it truly connected to better teaching?

Finally, what constitutes exemplary contributions to the department must be well defined. Do online courses require more work than teaching face to face? It probably depends on how often instructors participate in online discussions, how quickly they respond to e-mails, and whether feedback to students is prompt and meaningful.

All this gathering of information in addition to SRI demands time from already heavily scheduled faculty, and thoughtful analyses of the additional information demands yet more time of faculty and administrators. Yet given the key role of excellent teaching in the successes of institutions of higher learning, the time and effort devoted to making the best teachers possible is amply justified.

Weight Measures According to Reliability and Validity

Departments should decide in advance acceptable ranges of weights for the multiple measures. Those that have the highest reliability and validity should be given the greatest weight. For example, because SRI represent the perceptions of multiple raters across multiple occasions, they should be weighted more heavily than a single classroom observation by a colleague or administrator. Embedded assessments, by virtue of their direct measure of student learning, may also warrant significant weight.

Control for the Influence of Non-Instructional Variables

Is it true that SRI are never biased in any way? No. Teachers have legitimate concerns that circumstances beyond their control (a bias of some sort) and unrelated to how well they teach could negatively influence SRI. The National Council on Measurement in Education (NCME), in its website glossary, describes bias as "systematic errors in content, administration, and/or scoring that can cause test takers to get either lower or higher scores than their true ability would merit." The key is that the source of the bias is irrelevant to the trait being measured. Applied to student ratings, "Bias exists when a student, teacher, or course characteristic affects the evaluations made, either positively or negatively, *but is unrelated to any criteria of good teaching*" (Marsh, 2007, p. 350).

Using Marsh's definition, some non-teaching factors do, indeed, influence ratings but are not necessarily biases, because they may reflect either good teaching or engaged students. For example, IDEA SRI are positively correlated with students' desire to take the course (i.e., motivation), work habits, and class size (Benton et al., 2015). Students who are more motivated, apply themselves, and are enrolled in small classes *actually do tend to learn more*—an important criterion of effective teaching—and, therefore, tend to assign higher ratings. Nonetheless, institutions should control for the influence of these extraneous factors. The IDEA SRI system statistically attempts to level the playing field among faculty who teach courses of varying sizes and students of diverse motivations and work habits. For those who do not use IDEA's SRI, some attempt should be made to equitably allot teaching conditions to otherwise "level the playing field" by rotating assignments between required and elective courses, first-year and upper-level classes, and classes of small and large enrollments.

Provide Comparative Data

SRI tend to be negatively skewed—that is, scores are inflated. On a five-point scale, average global ratings of teaching excellence are typically around 4.0 (Benton et al., 2015). Moreover, they vary by academic disciplines—social science courses tend to earn higher ratings than math and "hard" sciences (Braskamp & Ory, 1994; Cashin, 1990; Centra, 1993, 2009; Hoyt & Lee, 2002b; Sixbury & Cashin, 1995). How, then, can one compare a score of 4.2 in an English class with a 3.8 in physics? The only fair comparison is to express the scores in standard deviation units from the respective discipline's mean. The 4.2 might be close to the average score in English, whereas the 3.8 could be above the average for physics. IDEA provides comparative scores by discipline to account for such differences. Institutions that do not use IDEA should examine whether SRI differences by discipline are attributable to variations in quality of teaching, student background preparation, or subject-matter difficulty.

Related to this is the finding that student self-ratings of their background preparation (Benton et al., 2015) and subject-matter difficulty (Centra, 1993, 2003; Marsh, 2001; Marsh & Roche, 2000) are correlated with SRI. Ratings tend to be higher in courses where students believe they had strong background preparation. Difficulty, on the other hand, shows a nonlinear relationship: Courses are rated lower when they are perceived as either too difficult or too elementary, a phenomenon known as the Goldilocks Effect. IDEA controls for these student variables via its adjusted scores. Institutions not using IDEA's SRI should include items that tap into students' perceptions of background preparation and course difficulty. Alternatively, teaching assignments may be rotated among courses that vary in student level and difficulty of material.

Include Global Items

Global or summary items provide evaluators a view of how students judged the overall quality of teaching and the course as well as how much they learned (Braskamp & Ory, 1994; Cashin, 1999; Centra, 1993). IDEA's overall ratings of the excellence of teaching and the course are highly correlated with average student ratings of progress on relevant course objectives (PRO). In turn, PRO is positively correlated with performance on course exams (Benton, Duchon, & Pallett, 2011). Global items provide a more valid measure of overall impressions than does averaging several dissimilar items together into a single score (Cashin, 1999).

Vary Evaluation Schedules

How frequently and for how many courses faculty should administer SRI ought to depend on faculty status and the purpose of the evaluation (Hoyt & Pallett, 1999). For first-year faculty it might make sense to collect student ratings data for every course and section. By the time employment recommendations are made, at least two sets of ratings should have been collected for the faculty member. The same holds true for decisions about promotions in rank. If teaching is connected to merit salary recommendations, all faculty must have some annual SRI data available. For tenured faculty, student ratings might be collected for each course biennially.

Use Written Comments Only Formatively

Although student written comments correlate positively with quantitative global ratings of the instructor (Braskamp, Ory, & Pieper, 1981; Burdsal & Harrison, 2008; Ory, Braskamp, & Pieper, 1980), faculty generally consider them less credible for decisions about tenure and promotion. Their chief value lies in the contributions they can make to improving teaching or the course (Braskamp et al., 1981), unless the institution employs sophisticated qualitative data analyses.

Employ Standardized Administration Procedures

Faculty must have confidence that ratings are collected similarly across courses and instructors. Written instructions to students should be standardized. Instructors should leave the room because ratings tend to be higher when the instructor is present (Braskamp & Ory, 1994; Centra, 1993; Feldman, 1979; Marsh & Dunkin, 1992). Ratings should be collected by a neutral party and the data taken to a location where they remain unavailable to the instructor until after grades have been submitted (Cashin, 1999). To ensure monitoring of procedures, students should be informed of policies and provided the means to report instructors who violate them.

Protect Student Confidentiality

Just as faculty must have confidence in the system, students must be assured their responses will remain confidential. Inform students that data will be held in a secure environment, will only be analyzed at the class level, and that results presented to the instructor will not be associated with any identifying information.

Encourage Good Response Rates

Concerns about low response rates have accompanied the increase in online survey administration. Although the online format may enhance standardization of administration procedures and better ensure student confidentiality, students are less likely to respond (Benton et al., 2015). Several constructive actions can be taken to increase the likelihood of a high response rate (http://ideaedu.org/support/existing-idea-paper-and-online-clients/idea-online-support/best-practices-for-online-response-rates). Instructors can *create value* for student feedback by placing IDEA relevant objectives alongside specific course objectives in the syllabus, informing students about modifications made in the course based on previous student feedback, encouraging them to complete the ratings, distributing a copy of a sample report given to instructors, and assuring confidentiality of responses. Institutions can *communicate* reminders through social media, university portals, learning management systems, department web sites, student publications, radio, flyers, and posters. For online surveys the instructors can ask students to bring a mobile device, tablet, or laptop to class on a day set aside for completing the ratings. Faculty should ensure the accuracy of students' "Respondent Identifier Labels." Ultimately, ongoing assessment should be championed as part of the institutional culture of monitoring and ensuring program quality.

Educate Administrators and Faculty

Given the ramifications of decisions made from teaching evaluations, administrators and faculty should be educated about how to interpret student ratings reports. It is imperative they understand that SRI should be part of a holistic evaluation of teaching. No major decision about a faculty member should ever be based on a single score. Evaluators should also recognize that all scores have error, and the estimated amount of error in reported scores should be communicated. Time should also be set aside to review, with the instructor, sample reports and interpret case studies based on student ratings data (Cashin, 1999). And administrators should have in mind remedies and/or specific suggestions for faculty; a plan of action for underperforming instructors should be designed and agreed to by both parties.

Conclusion

Validity (utility) is not a characteristic inherent in any measure, be it student ratings, classroom assessments, or standardized tests. Validity is tied to use—in how we interpret ratings and the actions we take based upon those interpretations. When used appropriately, SRI serve several important purposes. They can help faculty improve their teaching; administrators make decisions about salaries, reappointments, promotions, and tenure; institutions conduct program reviews; and students select courses. But, they should never be the only information source for such decision making.

Unfortunately, the valid use of ratings falls far short of its potential for several reasons (Pallett, 2006). Ratings tend to be overemphasized in summative decisions about teaching effectiveness. This overemphasis is easily explained. SRI can take little of a teacher's and department head's time compared to larger investments required in peer review, counseling, student interviews, accumulating other sources of teaching effectiveness, and so on. If a department is not seriously interested in improving teaching, SRI's inherent efficiencies tempt some departments into relying too much on them. Moreover, because SRI are quantifiable, they seem to lend themselves to widespread application; because SRI yield numbers, some people believe they can be applied across disciplines without adjustment, for example. Also, because SRI are efficient relative to other measures (e.g., peer and administrator observations, reviews of teaching materials), some institutions find it easy to forego the other measures entirely. And, of course, the most disappointing failure with SRI is that, even after faculty and department time has been spent on administering the SRI, faculty and departments do not capitalize on the investment they have already made; users often fail to reflect on SRI feedback and analyze the gathered information to make improvements. When ratings are overemphasized for summative decisions and underutilized for developmental purposes, their value is, unfortunately, reduced.

Our hope is that faculty, department heads, and other administrators who come across some of the SRI deniers' statements we referenced early on in this paper will be convinced by the overwhelming research regarding SRIs to deny the deniers' claims. SRI should count for something in a comprehensive faculty evaluation system. Student voices are critical, not only because they provide some quality control, but also because students have first-hand experience of what actually occurred in the classroom. When we take a comprehensive approach to evaluation of teaching, SRI should be one of the cornerstones of such evaluations no matter what the SRI deniers claim.

Steve Benton is Senior Research Officer at The IDEA Center where he leads a research team that designs and conducts reliability and validity studies for IDEA products. He is a Fellow in the American Psychological Association and American Educational Research Association, as well as an Emeritus Professor of Special Education, Counseling, and Student Affairs at Kansas State University where he served for 25 years. His current research focuses on effective use of student ratings of instruction for improving teaching and learning.

Ken Ryalls is President of The IDEA Center. After graduating from Indiana University with a PhD in Social Psychology, Ken has served in a variety of traditional roles in higher education, including professor, program director, division chair, dean, and vice president for academic affairs. He has always had an interest in teaching and learning, and spent his career at colleges dedicated to quality teaching with a student focus. He is the recipient of numerous teaching awards from both students and peers, and currently satisfies his passion for pedagogy through participation in research at IDEA, speaking and writing engagements both with clients and in the broader world of academia, and through development of new products and services that will allow faculty to improve their teaching performance.

References

Abrami, P. C., Perry, R. P., & Leventhal, L. (1982). The relationship between student personality characteristics, teacher ratings, and student achievement. *Journal of Educational Psychology, 74*, 111-125.

Adams, M. J., & Umbach, P. D. (2012). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education, 53*(5), 576-591.

Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 110-145). Beverly Hills, CA: Sage.

Allport, G. W. (1954). The nature of prejudice. Cambridge, MA: Perseus Books.

Amichai-Hamburger, Y., & McKenna, K. Y. (2006). The contact hypothesis reconsidered: Interacting via the internet. *Journal of Computer-Mediated Communication*, 11, 825–843. doi: 10.1111/j .1083-6101.2006.00037

Anderson, K. J., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Science, 27*, 184-201.

Arreola, R. A. (2006). *Developing a comprehensive faculty evaluation system* (2nd ed.). Bolton, MA: Anker Publishing.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.

Asher, L. (2013, October 27). When students rate teachers, standards drop. *The Wall Street Journal*. Retrieved from: http://www.wsj.com/news/articles/SB10001424052702304176904579115971990673400

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic SETs: Does an online delivery system influence student evaluations? *Journal of Economic Education, 37*, 21-37.

Basow, S. A., Codos, S., & Martin, J. L. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student Journal, 47*, 352-363.

Beleche, T., Fairris, D., & Marks, M. (2012). Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test. *Economics of Education Review, 31*(5), 709-719.

Benton, S. L., & Cashin, W. E. (2011). *IDEA Paper No. 50: Student ratings of teaching: A summary of research and literature*. Manhattan, KS: The IDEA Center. Retrieved from: http://ideaedu.org/wp-content/uploads/2014/11/idea-paper_50.pdf

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In Michael B. Paulsen (Ed.), *Higher Education: Handbook of Theory & Research*, Vol. 29 (pp. 279-326). Dordrecht, The Netherlands: Springer.

Benton, S. L., Duchon, D., & Pallett, W. H. (2013). Validity of self-report student ratings of instruction. *Assessment & Evaluation in Higher Education, 38*, 377-389.

Benton, S. L., Guo, M., Li, D., & Gross, A. (2013, April). *Student ratings, teacher standards, and critical thinking skills*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Benton, S. L., Li, D., Brown, R., Guo, M., & Sullivan, P. (2015). *IDEA Technical Report No. 18: Revising the IDEA Student Ratings of Instruction System*. Manhattan, KS: The IDEA Center. Retrieved from: http://ideaedu.org/wp-content/uploads/2015/12/Technical_report_18.pdf

Benton, S. L., Li, D., Brown, R., & Sullivan, P. (in press). *IDEA Technical Report No. 19: Analysis of IDEA student ratings of instruction system 2015 pilot data*. Manhattan, KS: The IDEA Center.

Benton, S. L., Webster, R., Gross, A. B., Pallett, W. (2010a). *IDEA Technical Report No. 15: An analysis of IDEA Student Ratings of Instruction in traditional versus online courses, 2002-2008 data*. Manhattan, KS: The IDEA Center. Retrieved from: http://ideaedu.org/wp-content/uploads/2014/11/techreport-15.pdf

Benton, S. L., Webster, R., Gross, A. B., Pallett, W. (2010b). *IDEA Technical Report No. 16: An analysis of IDEA Student Ratings of Instruction using paper versus online survey methods, 2002-2008 data*. Manhattan, KS: The IDEA Center. Retrieved from: http://ideaedu.org/wp-content/uploads/2014/11/techreport-16.pdf

Berrett, D. (2015, December 18). Scholars take aim at student evaluations' 'air of objectivity.' *The Chronicle of Higher Education*. Retrieved from: http://chronicle.com/article/Scholars-Take-Aim-at-Student/148859/

Boring, A., Ottoboni, K., & Stark, Ph. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *Science Open Research*. Retrieved from: https://www.scienceopen.com/document/vid/818d8ec0-5908-47d8-86b4-5dc38f04b23e

Braskamp, L. A, & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.

Braskamp, L. A., Ory, J. C., & Pieper, D. M. (1981). Student written comments: Dimensions of instructional quality. *Journal of Educational Psychology, 73*, 65-70.

Brinko, K. T. (1990). Instructional consultation with feedback in higher education. *Journal of Higher Education, 61*, 65-83.

Buchert, S., Laws, E. L., Epperson, J. M., & Bregman, N. J. (2008). First impressions and professor reputation: Influence on student evaluations of instruction. *Social Psychology of Education, 11*, 397-408.

Burbano, C. M. (1987). *The effects of different forms of student ratings feedback on subsequent student ratings of part-time faculty* (Doctoral dissertation). University of Florida, Gainesville.

Burdsal, C. A., & Harrison, P. D. (2008). Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. *Assessment and Evaluation in Higher Education, 33*, 567-576.

Cashin, W. E. (1990). Students do rate different academic fields differently. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice: New Directions for Teaching and Learning, No. 43* (pp. 113-121). San Francisco: Jossey-Bass.

Cashin, W. E. (1996). *Developing an Effective Faculty Evaluation System*. IDEA Paper No. 33. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development. Retrieved from: http://ideaedu.org/wp-content/uploads/2014/11/Idea_Paper_33.pdf

Cashin, W. E., (1999). Student ratings of teaching: Uses and misuses. In P. Seldin (Ed.), *Current practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions*. Bolton, MA: Anker.

Center of Inquiry (2013). *Wabash National Study 2006-2012*. Retrieved from: http://www.liberalarts.wabash.edu/study-overview/

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: Jossey-Bass.

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

Centra, J. A. (2009). *Differences in responses to the Student Instructional Report: Is it bias?* Princeton, NJ: Educational Testing Service.

Centra, J. A., & Gaubatz, N. B. (2000). Is there a gender bias in student evaluations of teaching? *Journal of Higher Education, 70*, 17-33.

Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education, 28(2)*, 149-160.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13*, 321-341.

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

Cohen, P. A. (1987, April). *A critical analysis and reanalysis of the multisection validity meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Cohen, P. A., & McKeachie, W. J. (1980). The role of colleagues in the evaluation of teaching. *Improving College and University Teaching, 28*, 147-154.

Cooper, T., & Terrell, T. (2013). *What are institutions spending on assessment? Is it worth the cost?* University of Illinois: National Institute for Learning Outcomes Assessment.

Costin, F. (1968). A graduate course in the teaching of psychology: Description and evaluation. *Journal of Teacher Education, 19*, 425-432.

Davis, B. G. (2009). *Tools for teaching* (2nd ed.). San Francisco: Jossey-Bass.

Dorn, D. S. (1987). The first day of class: Problems and strategies. *Teaching Sociology, 15*, 61-72.

Feeley, T. H. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education, 51(3)*, 225-236.

Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education, 4*, 69-111.

Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education, 10*, 149-172.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education, 18*, 3-124.

Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education, 24*, 129-213.

Feldman, K. A. (1987). Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration. *Research in Higher Education, 26*, 227-298.

Feldman, K. A. (1989a). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators and external (neutral) observers. *Research in Higher Education, 30*, 137-194.

Feldman, K. A. (1989b). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*, 583-645.

Feldman, K. A. (1992). College students' views of male and female college teachers: Part I Evidence from the social laboratory and experiments. *Research in Higher Education, 33*, 317-375.

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II Evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34*, 151-211.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93-129). Dordrecht, The Netherlands: Springer.

Finnegan, R. S. (2013). Linking teacher self-efficacy to teacher evaluations. *Journal of Cross-Disciplinary Perspectives in Education, 6*, 18-25.

Flaherty, C. (2014, May 30). A 'growth' field. *Inside Higher Education*. Retrieved from: https://www.insidehighered.com/news/2014/05/30/some-teaching-and-learning-centers-have-closed-after-recession-field-growing-over

Flaherty, C. (2016, January 11). Bias against female instructors. *Inside Higher Education*. Retrieved from: https://www.insidehighered.com/news/2016/01/11/new-analysis-offers-more-evidence-against-student-evaluationsteaching

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life.* New York, NY: The Free Press.

Gurung, R. A. R., & Vespia, K. J. (2007). Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology, 34(1)*, 5-10.

Halonen, J. S., Dunn, D. S., McCarthy, M. A., & Baker, S. C. (2012). Are you really above average? Documenting your teaching effectiveness. In R. A. R. Gurung & B. M. Schwartz, (Eds.), *Evidence-based teaching for higher education* (pp. 131-150). Washington, DC: American Psychological Association.

Hampton, S. E., & Reiser, R. A. (2004). Effects of a theory-based feedback and consultation process on instruction and learning in college classrooms. *Research in Higher Education, 45*, 497-527.

Hardy, N. (2003). Online ratings: Fact and fiction. In T. D. Johnson & D. L. Sorenson (Eds.), *Online student ratings of instruction: New Directions for Teaching and Learning*, No. 96 (pp. 31-38). San Francisco: Jossey-Bass.

Hareli, S., & Weiner, B. (2002). Social emotions and personality influences: A scaffold for a new direction in the study of achievement motivation. *Educational Psychologist, 37*, 183-193.

Hativa, N. (2013a). *Student ratings of instruction: A practical approach to designing, operating, and reporting*. Oron Publications.

Hativa, N. (2013b). *Student ratings of instruction: Recognizing effective teaching*. Oron Publications.

Howard, G. S., & Maxwell, S. E. (1980). The correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology, 72*, 810-820.

Howard, G. S., & Maxwell, S. E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education, 16*, 175-188.

Hoy, A. W. (2000). *Changes in teacher efficacy during the early years of teaching*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Hoyt, D. P., & Lee, E. (2002a). *Technical Report No. 12: Basic data for the revised IDEA system*. Manhattan, KS: The IDEA Center. Retrieved from: http://ideaedu.org/wp-content/uploads/2014/11/techreport-12.pdf

Hoyt, D. P., & Lee, E. (2002b). *Technical Report No. 13: Disciplinary differences in student ratings*. Manhattan, KS: The IDEA Center. Retrieved from: http://ideaedu.org/wp-content/uploads/2014/11/techreport-13.pdf

Hoyt, D. P., & Pallett, W. H. (1999). *IDEA Paper No. 36, Appraising teaching effectiveness: Beyond student ratings*. Manhattan, KS: The IDEA Center. Retrieved from: http://ideaedu.org/wp-content/uploads/2014/11/idea_paper_36.pdf

Huston, T. (2005). Research report: Race and gender bias in student evaluations of teaching. Retrieved from: http://sun.skidmore.union.edu/sunNET/ResourceFiles/Huston_Race_Gender_TeachingEvals.pdf

Jenkins, S. J., & Downs, E. (2001). Relationship between faculty personality and student evaluation of courses. *College Student Journal 35*(4), 636-640.

Johnson, T. D. (2003). Online student ratings: Will students respond? In T. D. Johnson & D. L. Sorenson (Eds.), *Online student ratings of instruction: New Directions for Teaching and Learning, No. 96* (pp. 49-59). San Francisco: Jossey-Bass.

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

Keig L., & Waggoner, M. D. (1994). Collaborative peer review: The role of faculty in improving college teaching. *ASHE-ERIC Higher Education Report No. 2*. Washington, DC: The George Washington University, School of Education and Human Development.

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Upper Saddle River, NJ: Prentice-Hall.

Knol, M. (2013). Improving university lectures with feedback and consultation. Academisch Proefschrift. Ipskamp Drukkers, B. V.

Kuh, G. D., & Hu, S. (2001). The effects of student-faculty interaction in the 1990s. *Review of Higher Education, 24*, 309-332.

Landrum, R. E. (2009). Are there instructional differences between full-time and part-time faculty? *College Teaching, 57*, 23-26.

Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction (electronic version). *Research in Higher Education, 40*(2), 221–232.

Leung, D. Y. P., & Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper through the Internet. *Research in Higher Education, 46*, 571-591.

Li, D., Benton, S. L., & Ryalls, K. (in press). *IDEA Research Report #10: Is there gender bias in IDEA Student Ratings of Instruction?* Manhattan, KS: The IDEA Center.

Linse, A. R. (2012). Faculty strategies for encouraging their students to fill out the SRTEs. Retrieved from: http://www.schreyerinstitute.psu.edu/IncreaseSRTERespRate/

Ludwig, J. M., & Meacham, J. A. (1997). Teaching controversial courses: Student evaluations of instructors and content. *Educational Research Quarterly, 21*, 27-38.

MacNell, L., Driscoll, A., & Hunt, A. N. (2014). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*. doi: 10.1007/s10755-014-9313-4

Manary, M. P., Boulding, W., Staelin, R., & Glickman, S. W. (2013). The patient experience and health outcomes. *The New England Journal of Medicine, 368*, 201-203.

Marincovich, M. (1999). Using student feedback to improve teaching. In P. Seldin, & Associates, *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/ tenure decisions (pp. 45-69)*. Bolton, MA: Anker.

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on student evaluations of teaching. *American Educational Research Journal, 38*, 183-212.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). Dordrecht, The Netherlands: Springer.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research, Vol. 8*. New York: Agathon Press.

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241-320). New York: Agathon Press.

Marsh, H. W., & Hattie, J. (2002). The relation between research productivity and teaching effectiveness. *Journal of Higher Education, 73*, 603-641.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching & Teacher Education, 7*, 303-314.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluation by their students. *Journal of Educational Psychology, 71*, 149-160.

Marsh, H. W., & Roche, L. A. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 30*, 217-251.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, and innocent bystanders. *Journal of Educational Psychology, 92*, 202-22.

McGhee, D. E., & Lowell, N. (2003). Psychometric properties of student ratings of instruction in online and on-campus courses. In T. D. Johnson & D. L. Sorenson (Eds.), *Online student ratings of instruction: New Directions for Teaching and Learning, No. 96* (pp. 39-48). San Francisco: Jossey-Bass.

McGowan, W. R., and Graham, C. R. (2009). Factors contributing to improved teaching performance. *Innovative Higher Education, 34*, 161-171.

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

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

Mulhere, K. (2014, December 10). Students praise male professors. *Inside Higher Education*. Retrieved from: https://www.insidehighered.com/news/2014/12/10/study-finds-gender-perception-affects-evaluations

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 75*, 138-149.

Nadler, M. K., Shore, C., Taylor, B. A. P., & Bakker, A. I. (2012). Making waves: Demonstrating a CTL'S impact on teaching and learning. *Journal on Centers for Teaching and Learning, 4*, 5-32.

Nilson, L. B. (2013). Time to raise questions about student ratings. In J. E. Groccia & L. Cruz (Eds.), *To improve the academy: Resources for faculty, instructional, and organizational development, Vol. 31*, (pp. 213-227). San Francisco, CA: Jossey-Bass.

Ormrod, J. E. (2014). *Educational psychology: Developing learners*. Upper Saddle River, NJ: Pearson Education.

Ory, J. C., Braskamp, L. A., & Pieper, D. M. (1980). Congruency of student evaluative information collected by three methods. *Journal of Educational Psychology, 72*, 181-185.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In T. D. Johnson & D. L. Sorenson (Eds.), *Online student ratings of instruction: New Directions for Teaching and Learning, No. 27.5* (pp. 27-44). San Francisco: Jossey-Bass.

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology, 72*, 321-325.

Palermo, J. (2013). Linking student evaluations to institutional goals: A change story. *Assessment & Evaluation in Higher Education, 38*, 211-223.

Pallett, W. H. (2006). Uses and abuses of student ratings. In P. Seldin, *Evaluating faculty performance* (pp. 50-65). Bolton, MA: Anker.

Patrick, C. L. (2011). Student evaluations of teaching: Effects of the Big Five personality traits, grades and validity hypothesis. *Assessment and Evaluation in Higher Education, 36*(2), 239-249.

Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: Meta-analysis. *Review of Educational Research, 74*, 215-253.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology, 90*, 751-783.

Porter, S. R., & Umbach, P. D. (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education, 47*, 229-247.

Porter, S. R., & Whitcomb, M. E. (2005). Non-response in student surveys: The role of demographics, engagement and personality. *Research in Higher Education, 46*, 127-152.

Quaglia, R. J., & Corso, M. J. (2014). *Student voice: The instrument of change*. Thousand Oaks, CA: Corwin.

Renaud, R. D., & Murray, H. G. (1996). Aging, personality, and teaching effectiveness in academic psychologists. *Research in Higher Education, 37*, 323-340.

Ryalls, K., Benton, S., Barr, J., & Li, D. (2016). Response to "Bias

Against Female Instructors" Editorial Note. Manhattan, KS: The IDEA Center. Retrieved from: http://ideaedu.org/research-and-papers/editorial-notes/response-to-bias-against-female-instructors

Shah, M., Nair, C. S., & Bennett, L. (2013). Factors influencing student choice to study at private higher education institutions. *Quality Assurance in Education, 21*, 402-416.

Sixbury, G. R., & Cashin, W. E. (1995). *IDEA Technical Report No. 10: Comparative data by academic field*. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Skaalvik, E. M., & Skaalvik, S. (2008). Teacher self-efficacy: Conceptual analysis and relations with teacher burnout and perceived school context. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *Self-processes, learning, and enabling human potential* (pp. 223-247). Charlotte, NC: Information Age.

Smith, B. P. (2007). Student ratings of teaching effectiveness: An analysis of end-of-course faculty evaluations. *College Student Journal, 41*, 788-800.

Smith, G., & Anderson, K. J. (2005). Students' ratings of professors: The teaching style contingency for Latino/a professors. *Journal of Latinos and Education, 4*, 115-136.

Smith, S. B., Smith, S. J., & Boone, R. (2000). Increasing access to teacher preparation: The effectiveness of traditional instructional methods in an online learning environment. *Journal of Special Education Technology, 15*(2), 37-46.

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. Published in *ScienceOpen, 1*, 1-26. Retrieved from: http://www.stat.berkeley.edu/~stark/Preprints/evaluations14.pdf. doi: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Sudkamp, A., Kaiser, J., & Moller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743-762.

Svinicki, M., & McKeachie, W. J. (2011). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (13th ed.). Belmont, CA: Wadsworth.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research, No. 109*, 45-56.

Thyer, B. A., Myers, L. L., & Nugent, W. R. Do regular social work faculty earn better student course evaluations than do adjunct faculty or doctoral students? *Journal of Teaching in Social Work, 31*, 365-377.

Twenge, J. (2006). *Generation Me*. New York, NY: Free Press (Simon & Schuster). ISBN 978-0-7432-7697-9

Venette, S., Sellnow, D., & McIntire, K. (2010). Charting new territory:

Assessing the online frontier of student ratings of instruction. *Assessment & Evaluation in Higher Education, 35*, 101-115.

Wang, A. Y., & Newlin, M. H. (2000). Characteristics of students who enroll and succeed in psychology web-based classes. *Journal of Educational Psychology, 92*, 137-143.

Ware, H., & Kitsantas, A. (2007). Teacher and collective efficacy beliefs as predictors of professional commitment. *Journal of Educational Research, 100*, 303-310.

Wieman, C. (2015). A better way to evaluate undergraduate teaching. *Change, Jan./Feb.*, 7-15.

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68-81.

Zakrajsek, T. (2006). Using evaluation data to improve teaching effectiveness. In P. Seldin, *Evaluating faculty performance: A practical guide to assessing teaching, research, and service* (pp. 166-180). Bolton, MA: Anker.

Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment and Evaluation in Higher Education, 37*, 227-235.

Zimmerman, J. (2014, January 24). The real scandal behind the Yale course Web site. *The Washington Post*. Retrieved from: https://www.washingtonpost.com/opinions/the-real-scandal-behind-the-yale-course-web-site/2014/01/24/f719ef56-8449-11e3-9dd4-e7278db80d86_story.html

T: 800.255.2757
T: 785.320.2400

301 South Fourth St., Suite 200
Manhattan, KS 66502-6209
E: info@IDEAedu.org
IDEAedu.org

Our research and publications, which benefit the higher education community, are supported by charitable contributions like yours. Please consider making a tax-deductible donation to IDEA to sustain our research now and into the future.

Notes