Chapter 7
Student Ratings of Instruction in College and University Courses

Stephen L. Benton and William E. Cashin

Introduction

The history of research on student ratings probably dates back to the 1920s, beginning with E. T. Guthrie's work at the University of Washington (Murray 2005). The topic has been studied more than any other in higher education, perhaps because of the role student ratings play in evaluating teaching and the availability of large datasets to analyze on most college campuses. The surge in studies began in the 1970s and continued into the 1990s. Whereas much of the student ratings research during those decades was published in higher education mainstream journals, many of the recent studies have appeared in discipline-specific publications (e.g., marketing, business). In addition, student ratings have found their way into the popular press. In some cases, authors have disparaged them and called for other measures of teaching effectiveness (e.g., value-added measures, peer ratings). They have written of the limitations in ratings—especially when used as the sole measure of teaching quality—often without making reference to the decades of research supporting their value when used appropriately.

In this chapter, we elaborate upon a previous publication: *IDEA Paper No. 50 Student Ratings of Teaching: A Summary of Research and Literature* (Benton and Cashin 2012). Some of the content in *IDEA Paper No. 50* is retained where no subsequently published study has changed its basic conclusions.[1] However, we include studies or reviews of the literature that provide questions, modifications, or

[1] Content contained in IDEA Paper No. 50 is reprinted by permission of The IDEA Center.

S.L. Benton, Ph.D. (✉)
The IDEA Center, 301 South Fourth St., Suite 200, Manhattan, KS 66502, USA
e-mail: steve@theideacenter.org

W.E. Cashin, Ph.D.
Kansas State University, Manhattan, KS, USA
e-mail: billcashin@md.metrocast.net

M.B. Paulsen (ed.), *Higher Education: Handbook of Theory and Research: Volume 29*,
DOI 10.1007/978-94-017-8005-6_7,

further support for its conclusions. New topics contained here but not covered in Benton and Cashin (2012) include the validity of self-reported data, issues to consider in reviewing student ratings research, correlating ratings with achievement in a subsequent course, and the effect of instructor first impressions. In addition, we elaborate upon and challenge the misconceptions that permeate discussions of student ratings.

We summarize the conclusions of major studies and reviews of the literature from the 1970s to 2013, a literature that is extensive and complex. We acknowledge, however, that this handbook chapter can offer only broad, general summaries and limited citations. At the end of 2012, there were 3,048 references in the ERIC database using the descriptor "student evaluation of teacher performance," the ERIC descriptor for student ratings of teaching/student evaluations of teaching (SRT/SET). By adding the descriptor "higher education," the number was reduced to 1,874. Restricting our search to the years 1994–2012 yielded 564 references. No major summary of the student ratings research was found in those 564 references, only specific studies. However, ERIC no longer includes chapters from the annual *Higher Education: Handbook of Theory and Research* or compilations of chapters from *Effective Teaching in Higher Education: Research and Practice* (Perry and Smart 1997) or *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (Perry and Smart 2007). We also searched PsycINFO, restricting the search to 2009–2012 publications in scholarly journals. This resulted in 329 hits, only 16 of which were relevant.

We found especially useful the following chapters published in the book by Perry and Smart (2007): Abrami et al. (2007), Feldman (2007), Marsh (2007), Murray (2007), and Theall and Feldman (2007). Those interested are encouraged to read these reviews as well as Hativa's (2013a, b) recent textbooks for more details. For readers with less time, Davis (2009), Forsyth (2003), Svinicki and McKeachie (2011), and Wachtel (1998), as well as earlier works by Braskamp and Ory (1994) and Centra (1993), have sections summarizing student ratings research.

Although the ERIC descriptor for student ratings is "student evaluation of teacher performance," we prefer the term *student ratings of instruction* (SRI). Whereas "evaluation" has a definitive and terminal connotation of determining worth, "ratings" refer to data that need interpretation. Using the term "rating" rather than "evaluation" helps to distinguish between the people who provide the information (sources of data) and those who interpret it (evaluators) in combination with other sources of information. Viewing student ratings as data rather than as evaluations puts them in their proper perspective.

Writers on faculty evaluation are almost universal in recommending the use of multiple sources of data. No single source of information—including student ratings—provides sufficient evidence to make a valid judgment about an instructor's overall teaching effectiveness. Further, there are important aspects of teaching that students are not competent to rate (e.g., subject-matter knowledge, course design, curriculum development, commitment to teaching, goals and content of the course, quality of student evaluation, support of department's accreditation

and program review requirements). For elaborations on these and other issues, we advise reading several sources (Abrami et al. 2007; Arreola 2006; Braskamp and Ory 1994; Cashin 1989, 2003; Centra 1993; Davis 2009; Forsyth 2003; Hativa 2013a, b; Marsh 2007; Svinicki and McKeachie 2011).

Although multiple sources of information are essential for making valid judgments of teaching effectiveness, no source of information is more reliable than student ratings, because they are based on multiple students who observe instruction on multiple occasions. Most higher education institutions, therefore, have some system in place for collecting student ratings data from classes on a systematic basis. This is because it is a relatively simple method for gauging student impressions of the course and the teacher's effectiveness. Moreover, students—who are the ones personally affected by instruction—have substantial opportunity to assess actual teaching behaviors, which enhances SRI validity. This may be why McKeachie (1979) calls student ratings the single most valid source of data collected on teaching effectiveness.

In spite of such advantages, considerable opposition to student ratings persists. One must acknowledge that some student ratings instruments are not well constructed and have little evidence of reliability or validity. Oftentimes, administrative procedures are not standardized, which prevents fair comparability among faculty. Some systems do not control for factors outside the instructor's control that affect student ratings, such as student motivation and work habits, and class size (Hoyt and Pallett n.d.). Most troublesome is the practice of making student ratings the only source of evidence for assessing teaching effectiveness. Multiple sources of information should always be collected to triangulate evidence (Berk 2005; Hativa 2013b; Hobson and Talbot 2001).

Validity of Self-Reported Data

Both authors have made numerous presentations about student ratings on college and university campuses where they have sometimes encountered skepticism from some faculty about whether self-reported student ratings can be trusted. Researchers across a variety of behavioral and social sciences have investigated the accuracy of self-reported data (e.g., Brener et al. 2003; Costin et al. 1971; Del Boca and Noll 2002; Kuncel et al. 2005). A review of the literature reveals that, in general, self-report can be interpreted validly under certain conditions (Babor et al. 2000; Babor and Del Boca 1992; Patrick et al. 1994; Freier et al. 1991; Midanik 1988; Cooper et al. 1981). First, validity depends on the sensitivity of the information being sought. For example, asking about course assignments is less sensitive than asking about the overall quality of teaching. The sensitivity of student ratings depends on the instructor's and institution's commitment to confidentiality, the reporting of only class averages, and the assurance the instructor cannot view the class report until after final grades are submitted. Second, the validity of self-report can

be enhanced when paired with other forms of evidence. Additional indicators of teaching effectiveness (e.g., peer ratings, student learning outcomes) should be collected to confirm the validity of student ratings, especially when the information is sensitive. Third, preparing students to complete the ratings can help to improve validity. The instructor should take time to encourage students to take the process seriously. Students should be informed that the ratings are used to make important decisions about the course. The instructor should administer ratings when students are alert, engaged, and have adequate time to complete the process. Fourth, instruments with multiple items usually have more validity than those with only a few items, because teaching effectiveness is a complex construct. Students' responses to many questions should be collected to obtain a complete picture of their impressions of the teaching and learning experience. Moreover, ratings aggregated across students lead to greater reliability (Ray 1987).

Misconceptions About Student Ratings

In spite of the evidence supporting the reliability and validity of student ratings, misconceptions persist. According to cognitive psychologists, people develop misconceptions for several reasons. Sometimes they arise out of confirmation bias—the tendency to look for information that confirms our existing beliefs and ignore or discredit information that contradicts them. Instructors who focus on the occasional negative written comment or a lower than expected overall rating of the course may overlook the formative value of student feedback. Another possibility is that we may hold a personal or emotional investment in a belief (e.g., that we are an above average teacher). If we get average or low ratings, then there must be something wrong with either the students or the student ratings instrument. In addition, people can hold misconceptions because their beliefs are consistent with their experience. For example, some instructors perceive pressure from students to assign high grades. At the same time, they may feel pressure within their department to receive high student ratings. The juxtaposition of those two feelings could lead one to make a cause-effect connection. Another reason an individual might hold on to a misconception is because it follows from a general worldview. If someone believes, for example, that students are just not capable of being objective, then that colors his or her faith in the interpretability of ratings.

Regardless of why misconceptions endure, they should be recognized for what they are—incorrect or partial information. Several authors have brought to light misconceptions about student ratings that are unsupported by research and that make improved practice difficult (Aleamoni 1987; Feldman 2007; Kulik 2001; Svinicki and McKeachie 2011; Theall and Feldman 2007). What follows are some of the more common misconceptions about student ratings and what the research literature generally says about them. In later sections of this chapter, we address most of them more specifically.

Students Are Not Competent Enough to Rate Teaching

Students observe classroom teaching behaviors more than anyone else. For that reason alone, it makes sense to consider what they have to say. In the section on "Validity," we summarize research that finds student ratings correlate positively with other measures of teaching effectiveness. Examples include student achievement measures (e.g., final exam scores, standardized tests); motivation for future learning and attitude change; the teacher's self-ratings of multiple sections of the same course; and ratings by administrators, colleagues, alumni, and trained observers.

Students Cannot Make Consistent Judgments

Students are very consistent in their ratings of teacher behaviors, their own learning, and of overall impressions of the course and teacher. In the section on "Reliability," we review evidence that shows agreement among students rating the same instructor/course, stability in ratings of the same instructors across time, and generalizability of ratings across the same course and different courses taught by the same instructor.

Student Ratings Are Just Popularity Contests

The effects of teacher personality traits (e.g., enthusiasm) on student ratings are weak. The effects are most likely explained by the behaviors the instructor exhibits in class rather than by who they are as a person. This research (e.g., Braskamp and Ory 1994; Centra 1993) is found in the section on instructor variables not related to student ratings. Student ratings instruments that are backed by reliability and validity evidence typically assess more than what students think of the teacher. Students rate teaching methods, how much they have learned, various aspects of the course, quality of teaching, and their overall impressions of the course.

Students Will Not Appreciate Good Teaching Until They Are Out of College a Few Years

As addressed in the section on correlating student ratings with other criteria, alumni retrospective ratings of instructors correlate positively with those given by current students.

Students Just Want Easy Courses

Ratings are actually higher when students report the instructor sets high achievement standards for the course. The relationship between ratings and course difficulty is

curvilinear. Ratings tend to be lower when the class is either too easy or too difficult and highest when the class is appropriately challenging. The solution is to discover what is "just right." We review much of the research behind these findings in the section on student variables that correlate with ratings.

Student Feedback Cannot Be Used to Help Improve Instruction

Combining consultation with feedback from student ratings is more useful for improving instruction than providing feedback alone. The greatest improvement comes when feedback and consultation target teacher behaviors that address problems students identify. We review this research in the section on "Usefulness of Student Ratings."

Emphasis on Student Ratings Has Led to Grade Inflation

The grade students expect to receive in a course actually has little effect on student ratings. Even when low correlations are observed, it may not necessarily indicate grading leniency on the part of the instructor. It could be that students who learn more earn higher grades and, therefore, give higher ratings. Or student characteristics, such as interest and motivation, may lead to more learning, higher grades, and better ratings. We review this research in the section on "Student Variables Related to Student Ratings That May Require Control."

Because Students of the "Millennial Generation" Are Different from Previous Generations, Student Ratings Are no Longer Valid

Nilson (2013) has argued that students of the "millennial generation" feel entitled to receiving high grades without putting out much effort. The implication is that contemporary students award low ratings to instructors who have high standards and grade strictly. Actually, the evidence is very clear that across the years, students have not changed markedly in their perceptions of teaching, the course, or their learning. Ratings on average tend to be high (Benton et al. 2010a, b, 2012a; Hoyt and Lee 2002a). Moreover, contrary to what some might expect, students rate the quality of the instructor and course higher when they perceive the instructor has high standards and expects them to share responsibility for their own learning (Benton et al. 2013). Students also assign higher ratings of teaching effectiveness when the instructor encourages them to think for themselves (Zhao and Gallant 2012).

Such misconceptions about student ratings ignore more than 50 years of credible research. They persist, unfortunately, largely due to ignorance of the research, personal biases, suspicion, fear, and general hostility toward any evaluation process (Theall and Feldman 2007).

Issues to Consider When Evaluating Student Ratings Research

With the recent increase in studies investigating student ratings, some care should be exercised in assessing their quality. Numerous issues should be considered in evaluating the credibility of any study, and we refer interested readers to a couple of classic sources (Campbell and Stanley 1963; Krathwohl 1998). A key consideration is whether the researcher has offered a *credible rationale* for the design of the study. For example, does the tone of the article suggest a bias either for or against student ratings? If so, then the reader must carefully scrutinize the article by reading details about the methods, analyses, and conclusions (which one should do anyway). Related to this is whether the explanation or rationale for the study is built upon previous research and thinking in the field. A reference, for example, to the "Dr. Fox" effect—where a professional actor who delivered a dramatic lecture but with little meaningful content received high ratings—without also citing research critical of that study would show a failure to examine previous thinking on the subject (see section on "Expressiveness" under "Instructor Variables Related to Student Ratings That May Require Control").

The reader should also examine *how credibly the study was conducted*. If data were collected from college students, how representative were they of most students? Did the study take place in a single classroom or across multiple classrooms? Was evidence provided regarding the reliability and validity of the measures used in the study? If student learning/achievement was an outcome variable, how was it defined and measured? Also of concern is whether one could reasonably replicate the procedures in a typical college setting. For example, some campuses have more control than others over how students are assigned to different course sections.

A third consideration is the *credibility of the empirical evidence* provided. Is there cause to question the authenticity of the evidence? For example, were student ratings administered in the presence of the instructor? (This is something not recommended—see section on "Manipulating Administrative Procedures.") If so, how might this have affected the results?

The reader should also determine whether the author has reasonably attempted to address or eliminate rival explanations for the findings. Although researchers try to control extraneous variables, there are usually alternative explanations for the outcome beyond that which was hypothesized. A careful researcher ponders rival explanations in advance and then designs the study in a way to make them less plausible.

In the sections that follow, we briefly summarize research that provides evidence of the reliability and validity of student ratings of instruction. We also examine extraneous student, instructor, and course characteristics that are either unrelated or related to ratings. Finally, we review research that compares ratings administered online versus on paper and between courses conducted online versus face-to-face. We conclude by discussing the usefulness of student ratings.

Reliability

Reliability refers to the consistency, stability, and generalizability of student ratings data. Of primary concern is the *consistency* in ratings among students in the same class rating the same instructor. Reliability coefficients typically range from .00 to 1.00 with higher values indicating greater consistency. Standard errors of measurement (SEM), which should be reported along with reliability coefficients, indicate the amount of error or spread (+ or −) in the scores. Reliability estimates vary, depending upon the number of students completing the ratings. As the number of students responding increases, reliability (or consistency in the scores) increases, and the amount of spread in the scores (error) decreases.

Researchers have found consistency in ratings completed by students in the same class. Average split-half reliability coefficients for the *Individual Development and Educational Assessment* (IDEA) student ratings items range from .78 for small class sizes (10–14 students) to .94 for classes with 50 or more students (Hoyt and Lee 2002a). Cronbach alpha internal consistency coefficients for the five IDEA teaching style subscales range from .84 to .94. Marsh (1984) reported average coefficients for the *Student Evaluation of Educational Quality* (SEEQ) ranging from .74 for classrooms with 10 students to .95 for those with 50. The greater the number of students completing the ratings, the higher the reliability; lengthier instruments generally tend to have higher reliability than shorter ones.

Stability pertains to the amount of agreement in ratings across time. Ratings of the same instructor across semesters tend to be similar (Braskamp and Ory 1994; Centra 1993). Marsh and Hocevar (1991) conducted a longitudinal study, using the SEEQ. They compared an average of 30 ratings each for 195 teachers across 13 years and found almost no systematic changes across time among those with little, intermediate, or substantial teaching experience.

Generalizability refers to how reliably ratings reflect the instructor's general teaching effectiveness, not just how effectively he or she taught a particular course in a given term. Marsh and Bailey (1993) employed multivariate profile analysis on the SEEQ to develop unique profiles for each teacher evaluated across a 13-year period. The unique teacher profiles of SEEQ factors were consistent across courses. Hativa and Raviv (1996) found similar results in a sample of physics and chemistry instructors who administered ratings eight times across a 2-year period. The authors identified a teaching profile for each instructor that was stable across time.

Reliability is especially relevant when making personnel decisions about an instructor's general teaching effectiveness. Administrators should consult additional information beyond student ratings when making such decisions (see Cashin 1996, 2003). If the instructor teaches only one course (e.g., part-time instructors), then consistent ratings from two different terms may be sufficient. For most instructors, however, ratings from a variety of courses are necessary, preferably two or more courses from every term, for at least 2 years, totaling six to eight courses. If there are fewer than 10 raters per class, we recommend collecting data from 10 to 12 classes.

Validity

As with reliability, validity is not a characteristic inherent in a student ratings instrument. Validity is determined by how the ratings are used—how they are interpreted and what actions follow from those interpretations—referred to as the consequential basis of validity (Messick 1989). Notably, McKeachie (1997) cautioned that faculty and administrators need to be educated about how to use ratings appropriately (i.e., validly).

Student ratings typically serve several valid functions. They help faculty to improve their teaching and courses, help administrators make decisions about salary and promotion, help committee members select recipients of teaching awards, help institutions conduct program reviews, and help students select courses. When used in combination with other measures of teaching effectiveness, ratings can support all of these functions. However, when misused for other purposes (e.g., to base course content on student ratings form content, to make administrative decisions based on ultrafine discriminations in ratings), validity is threatened (Ory and Ryan 2001).

Researchers traditionally take one of several approaches to investigating the validity of student ratings. They correlate ratings from multiple sections of the same course with student achievement on a common examination; correlate ratings with other criteria (e.g., alumni ratings, peer ratings, or self-ratings); examine bias by correlating ratings with student, instructor, and course characteristics; manipulate administrative procedures; conduct experiments in nonnatural settings; and analyze the underlying dimensions of ratings (Ory and Ryan 2001). Evidence from all such studies supports *construct* validity (Messick 1995), which enables meaningful interpretation of the data collected.

In educational measurement, the basic question related to validity is: Does the test—the variable—measure what it is supposed to measure? For student ratings, this translates into: To what extent do student ratings items measure some aspect of teaching effectiveness? Unfortunately, there is no agreed-upon definition of what student ratings measure or any single, all-embracing criterion of effective teaching (see, e.g., Cashin 2003; Hativa 2013b). However, across multiple factor-analytic studies (Braskamp and Ory 1994; Feldman 1989a; Hativa et al. 2001; Marsh 1987; Murray 1997), the perceived teacher behaviors most highly correlated with effective instruction include organization, clarity, enthusiasm/expressiveness, and rapport/interactions (summarized in Hativa 2013a). In examining the items from student ratings instruments used in those studies, the following elements of effective teaching emerge:

1. Organization—course materials and teacher well prepared; lessons linked to overall course framework
2. Clarity—simplified explanations; understandable; links to students' prior knowledge
3. Enthusiasm/expression—being enthusiastic about the subject or about teaching; making dynamic presentations; using humor
4. Rapport/interactions—encouraging students to ask questions, to discuss and share ideas, and to invite a variety of viewpoints

Although these teacher characteristics are associated with effective teaching, researchers should continue to pursue various approaches to examining the validity of student ratings. They can do so by collecting data that either support or contest the conclusion that student ratings reflect effective teaching. In the paragraphs that follow, we summarize research to date employing these various approaches.

Correlating Student Ratings with Achievement in the Current Course

Perhaps the best indicator of effective teaching is student learning. Other things being equal, the students of more effective teachers should learn more. A number of researchers have attempted to examine this hypothesis by correlating student achievement with ratings across multi-section courses. The instructors use the same syllabus and textbook, in some cases the same assignments, and, most importantly, the same *external* final exam (i.e., an exam developed by someone *other* than the instructors). Student ratings of the course and instructor are then correlated with final exam scores. For example, Beleche and colleagues (Beleche et al. 2012) correlated students' ratings with performance on a final exam common to multiple sections of a remedial college course. The points students received on the measure of their learning correlated positively with global ratings of the overall excellence of the course.

Not only are global course ratings related to student achievement, but student ratings of their own learning also correlate positively with their actual performance. Benton and colleagues (Benton et al. 2011) examined student ratings in multiple sections of the same course taught by the same instructor. They correlated students' ratings of progress on objectives the instructor identified as relevant to the course with their performance on exams tied to those objectives. The instructor selected 2 of 12 learning objectives as relevant, using the IDEA *Faculty Information Form*. Students rated their progress on all 12 of the IDEA learning objectives. All exam items were tied to content related to the two relevant objectives selected for the course: "gaining factual knowledge (terminology, classifications, methods, trends)" and "learning fundamental principles, generalizations, or theories." The chair of the department, not the instructor, administered the ratings at the end of the semester. Student ratings of progress made on the two relevant objectives were positively correlated with their exam performance on content connected with the objectives ($r = .32$ and .33). In contrast, the average correlations between student ratings of the 10 other less relevant objectives and exam performance were negligible (average $r = -.03$).

Cohen (1981, 1987) and Feldman (1989b) reviewed several multi-section studies and, for each one, correlated final exam scores with various student ratings items.[2]

[2] The authors converted various summary statistics reported in the multi-section studies into Pearson product-moment correlations.

Table 7.1 Correlations between student final exam performance and various dimensions of student ratings

	Average correlations with final exam across three studies		
Student ratings of:	Cohen (1981)	Cohen (1987)	Feldman (1989b)
Achievement/learning	.47	.39	.46
Overall course	.47	.49	–
Overall instructor	.44	.45	–
Teacher skill	.50	.50	–
Course preparation	–	–	.57
Clarity of objectives	–	–	.35
Teacher structure	.47	.55	–
Understandableness	–	–	.56
Teacher rapport	.31	.32	–
Availability	–	–	.36
Respect for students	–	–	.23
Teacher interaction	.22	.52	–
Encouraging discussion	–	–	.36
Evaluation	–	.30	–
Feedback	–	.28	–
Interest/motivation	–	.15	–
Difficulty	–	−.04	

Table 7.1 presents the average correlations as they were reported in Cohen (1981, 1987) and Feldman (1989b). Both researchers identified the instructional dimensions (e.g., teacher preparation and course organization, teacher clarity, teacher stimulation of student interest, and students' perceived impact or outcome of the course) most highly correlated with student achievement. (See also Abrami 2001, and Kulik 2001, for support of the relationship between student learning and student ratings.)

In a follow-up study, Feldman (2007) reported the average correlations between a measure of student achievement and 24 specific instructional dimensions frequently measured by student ratings instruments. In a separate table, he also compared the correlations of various instructional dimensions with *student achievement* and *students' overall evaluation of the teacher*. The correlations with achievement and overall evaluations of teaching were not always of the same magnitude (e.g., quality and frequency of feedback correlated only .23 with student achievement but .87 with overall evaluation), but they showed the positive contribution of various instructional dimensions to the measure of their learning.

Given the restricted range in most student ratings scales, and the less than perfect reliability of classroom exams, the correlations reported in this section are impressive. Moreover, because teachers are not the only cause of student learning, and probably not the most important one, one would not expect students' ratings of instruction to correlate perfectly with how much they learn in a course (Hativa 2013b). Even so, the multi-section studies consistently show that classes in which the students gave the instructor higher ratings were the ones where the students learned more (i.e., scored higher on the exam).

Moreover, the correlations reported here are within the range of correlations found between *teachers' judgments* of students' achievement and students' actual achievement (see Sudkamp et al. 2012 for a meta-analysis of 75 studies). Although Sudkamp et al. reviewed studies conducted only with secondary teachers, one might predict that college and university instructors—who typically face larger enrollments, have less frequent student contact, and typically have less preparation than secondary teachers in how to create valid and reliable assessments—would make even poorer predictions of their students' achievement. We hypothesize that the accuracy of their predictions of how well their students would perform on achievement tests would be no better students' own estimations of how much they have learned.

In spite of the evidence that student ratings correlate with achievement, several authors have reported little or no relationship. For example, Galbraith and colleagues (Galbraith et al. 2012) conducted a multi-section analysis that examined correlations between a business school's Standardized Learning Outcome Assessment Test (SLOAT) and average student ratings on two measures. According to the authors, "The School's SLOAT exams are developed individually for each course in each program by a committee of content experts in the subject area…" (p. 358). So, unlike the studies reviewed thus far in this section, the SLOAT was not a common exam. The second measure, which they called "Course," was an average of eight items taken from the Student Perception of Teaching Effectiveness (SETE). However, the authors provided no information about the reliability and validity of either of these measures. Without such information, the meaning behind the low correlations reported between the SETE measures and the SLOAT is unclear. Moreover, because the course section was the unit of analysis, the sample size for each analysis was extremely small (ranged from 8 to 13 sections). When the authors conducted the analysis at the pooled level (i.e., putting all sections together), the sample size increased and the correlations were low and positive, which is similar to Feldman's (1989b) and Cohen's (1981, 1987) findings.

Similarly, McCarthy and colleagues (McCarthy et al. 2011) contend that student ratings do not provide evidence of student learning. Notably, the authors made no reference to Cohen (1981, 1987) or Feldman (1989b). They recommend that "embedded assessments"—analysis of formative measures of student learning (e.g., exams, papers)—are better measures of learning than student ratings, an idea long ago strongly recommended by Centra (1979, 1993), Braskamp and Ory (1994), Arreola (2006), and Cashin (1989, 2003).

In a meta-analysis of studies of the relationship between SRIs and measures of learning, Clayson (2009) reviewed 17 articles with 42 datasets, containing 1,115 course sections. He found, in general, a small positive association between measures of learning and SRIs, which is consistent with previous research. He concluded that the association between learning and SRIs is valid "to the extent that the student's perception of learning is valid" (p. 27). We agree. We would also argue that the learning/SRI association is more valid and stronger to the extent that the instructor's measure of learning is reliable and valid. This may be why he found higher correlations reported in studies published in education/psychology journals. Unlike their colleagues in other colleges, researchers in education and psychology

have typically been educated in educational/psychological measurement, which includes knowledge of how to construct reliable and valid tests. When the reliability of the assessment used to measure learning is higher, the correlation between learning and SRIs is higher, other things being equal.

Correlating Ratings with Achievement in a Subsequent Course

Another way of testing validity is to correlate student ratings in a prior course with grades in a subsequent course. In general, researchers have found only weak correlations (e.g., Beleche et al. 2012; Carrell and West 2010; Weinberg et al. 2009; Yunker and Yunker 2003). This value-added approach to studying the validity of student ratings makes several assumptions we find problematic. First, it assumes that grades in the subsequent course serve as a proxy for teaching effectiveness in the prior course. The argument is that the better the grade in the second course, the better was the teaching in the preceding course. However, this seems to ignore the possibility that grades in a subsequent class are subject to selection bias, because they do not account for students who drop out or choose not to take the next course. Second, it assumes that the subsequent course significantly builds upon the knowledge learned in the prior course—that the prior course lays a basic, necessary foundation for future learning. But if the subsequent course fails to augment the knowledge gained in the preceding course, the relationship between ratings and subsequent grades should be weak (Beleche et al. 2012). A third assumption is that the value added by an instructor depends, in part, on how well a subsequent instructor teaches. Since student ratings instruments are designed to assess student perceptions of teaching effectiveness in the *current* course, it defies logic as to why they should be correlated with outcomes of teaching in a *subsequent* course. Not even grades in a previous course are a good predictor of performance in a subsequent course (Grant 2007).

The challenges in implementing this approach are daunting. First, students must be randomly assigned to classes to control for selection bias that occurs from reasons students give for choosing one course over another. Upper-level students, for example, are more likely than first-year students to select course sections based on instructor reputation (Leventhal et al. 1975), which is moderately related to student ratings (Perry et al. 1974, 1979b). On the other hand, first-year students are far more likely to make section-selection decisions on the basis of time of day the class meets, which is unrelated to ratings (Aleamoni 1981; Feldman 1978). Even with random assignment, low-performing students might be less likely than high-performing students to advance to the subsequent course, which would still create selection bias. Second, analyses have to be conducted of student standardized test scores to demonstrate a level of confidence in the random assignments. Third, faculty who teach the same course have to use an

identical syllabus and give the same exams during a common testing period. Finally, instructors teaching the same course have to jointly grade all exams or have them scored by a common grader.

Sullivan and Skanes (1974) did carry out such a quasi-experimental design by randomly assigning first-year students to different sections of undergraduate courses. Across multiple disciplines and courses, the authors found modest but significant positive correlations between global ratings of the instructor's competence and student achievement in the current course. The authors then examined a sample of students enrolled in sections of introductory psychology who enrolled in a subsequent course (not randomly assigned) in psychology. The proportion of students who enrolled in the subsequent course was similar between those taught by instructors with high final exam scores and those taught by instructors with high ratings.

The research reviewed in the previous section provides evidence that student ratings are a valid—albeit not perfect—proxy for learning in the current course. However, in and of themselves, they are only weakly related to achievement in a subsequent course. Yet, Knol (2013) found that when the feedback from ratings is combined with consultation, students subsequently report greater learning from the instructor. Future research might examine whether feedback plus consultation could also result in greater student achievement in a subsequent course.

Correlating Ratings with Other Criteria

Student ratings are positively correlated with a number of other variables used to measure teaching effectiveness: instructor self-ratings; ratings by administrators, colleagues, alumni, and trained observers; and student written comments.

Instructor Self-Ratings

One correlate of student ratings that would presumably be acceptable to most faculty is self-ratings completed by the instructor. In a review of 19 studies, Feldman (1989a) reported an average correlation of .29 between student ratings and instructor self-ratings, often using the same instrument. In another study (Marsh et al. 1979), instructors were asked to rate their teaching effectiveness in two courses in order to see if the course the instructor rated highest also had the highest student ratings. Student ratings were indeed higher in the courses the instructors believed were taught more effectively. The median correlation—across six factor scores—was .49 between the instructor and student ratings. In a related study, Marsh (1982) found that 34 of the 35 correlations between student ratings and instructor self-ratings were statistically significant, with a median correlation of .30. Subsequently,

Marsh and Dunkin (1997) found a median correlation of .45 between instructor self-ratings and student ratings on nine scale scores. Such findings support the criterion-related validity of student ratings.

In spite of the consistent positive correlations reported between student ratings and instructor self-ratings, Feldman (1988) wanted to know whether students actually have an appropriate view of effective teaching. To address this concern, he reviewed 31 studies and found that students' views of effective teaching were very similar to the instructor's views (the average correlation was .71). However, he found some subtle differences. Students tended to assign more importance to the instructor being interesting, having good speaking skills, and being available to students. Students also focused more on the outcomes of instruction (e.g., what they learned). In contrast, instructors placed relatively more emphasis on challenging and motivating students, setting high standards, and fostering student self-initiated learning.

Feldman's (1988) findings might seem to imply that students undervalue the importance of setting high standards and fostering student self-initiated learning. However, that conclusion is refuted by other research. Using IDEA student ratings, Hornbeak (2009) found that students' desire to take a course from an instructor was positively correlated (.54) with how much the instructor expected students to take their share of responsibility for learning. Likewise, students have a stronger desire to take a course (.52) when they perceive the instructor has high achievement standards (Hoyt and Lee 2002a). In addition, student ratings of the instructor's achievement standards and expectations that students share in responsibility for learning are positively correlated with student self-ratings of learning and overall ratings of the course and instructor (Benton et al. 2013). So, although students may place relatively less value than instructors on standards and self-initiated learning, they report more progress and assign higher ratings when instructors demonstrate those characteristics.

Ratings by Administrators and Colleagues

Student ratings correlate moderately with administrator ratings of the instructor's general reputation. Coefficients range from .47 to .62 (Kulik and McKeachie 1975). Feldman (1989a), using global items, found an average correlation of .39 across 11 different studies.

Instructor ratings by colleagues that are *not* based on classroom observations are moderately correlated with student ratings, with correlations ranging from .48 to .69 (Kulik and McKeachie 1975). Feldman (1989a) found an average correlation of .55, using global ratings. However, ratings by colleagues can be unreliable and uncorrelated with student ratings when made by untrained observers during single classroom visitations where an unsystematic approach is employed, i.e., different faculty visiting the same class tend to disagree (Marsh 2007; Marsh and Dunkin 1997).

Ratings by Alumni

Some faculty may question whether current students can adequately judge the long-term effects of instruction. They may argue that students cannot truly appreciate how much they have learned or the value of a course until after they have graduated. However, current and former students tend to agree in their ratings, as average correlations typically range from .54 to .80 (Braskamp and Ory 1994). Overall and Marsh (1980) and Feldman (1989a), reviewing six studies, reported average correlations of .83 and .69, respectively. These findings belie the conventional wisdom that students only come to appreciate effective teaching *after* they graduate and enter into the real world as working adults.

Ratings by Trained Observers

A few studies examined the relationship between student ratings and external observers who were trained to make classroom observations (see Feldman 1989a; also Marsh and Dunkin 1997). Reviewing five studies, Feldman reported an average correlation of .50 between the ratings of trained observers and global student ratings. In a related study, Murray (1983) reported a median reliability of .76 among ratings by trained observers, which suggests ratings by colleagues might be more reliable if faculty were trained prior to making classroom observations.

Student Written Comments

The quantitative aggregated data provided from student ratings scales considerably overlaps the information contained in student written comments. In one study of 14 classes, Ory and colleagues (Ory et al. 1980) found a correlation of .93 between a global instructor item and students' written comments. In a second study of 60 classes, the authors (Braskamp et al. 1981) found a correlation of .75. More recently, Burdsal and Harrison (2008) reported a correlation of .79 in a sample of 208 classes. In a study of 80 courses at a single institution, Hativa (2013b) found a positive correlation between the percentage of positive student comments and ratings of overall teaching effectiveness.

These studies suggest that, for *personnel decisions*, student written comments usually reflect the information derived from quantitative scales. Nonetheless, when decisions are made about promotion, instructors generally regard written comments as less credible than student responses to objective questions. On the other hand, faculty rate written comments as more credible when the purpose is for self-improvement (Braskamp et al. 1981). We agree with others (Abrami 2001; Hativa 2013b; Marsh and Dunkin 1997) that written comments are more appropriate for formative than summative evaluation of teaching effectiveness.

The studies cited thus far provide evidence that student ratings are related to other measures of teaching effectiveness. Student ratings are positively related to student achievement in the current course, teacher self-ratings, administrator and colleague ratings, alumni ratings, ratings by trained observers, and student written comments. In the next section, we consider possible biases in student ratings.

Evaluating Possible Sources of Bias in Student Ratings

Faculty and administrators are sometimes concerned about possible biases in student ratings, such as class size and student interest in the course. Some writers have suggested that bias can be defined as anything *not under the control of the instructor*. However, Marsh (2007) offered another definition we find more valid: "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*" (p. 350; see also Centra 2003, p. 498). By this definition, the correlations between student ratings and class size, or between student ratings and student interest in the course, are not biases because students in small classes and students who are interested in the subject matter *actually do tend to learn* more and, hence, give their teachers higher ratings. Rather than using the term "bias," we distinguish between variables (when correlated with student ratings) that possibly require control and those that do not require control, especially when making personnel decisions.

Variables Related to Student Ratings That Are Not *Biases*

Despite widespread faculty concern, researchers have discovered relatively few variables that correlate with student ratings *that are not* also related to instructional effectiveness (i.e., student learning). Variables related to student ratings that are not biases include class size (ratings are somewhat higher in smaller classes), course level (ratings tend to be higher in higher-level courses), faculty rank (ratings are typically higher for higher-ranked faculty) (Feldman 2007), reason for taking the course (ratings are higher for elective compared to required courses and for students in their major) (Leventhal et al. 1975), and faculty reputation as an instructor (ratings are higher for faculty with good reputations) (Perry et al. 1974, 1979a, b).

Feldman discussed some factors that might account for why these variables should *not* be considered biases. For example, at certain institutions, higher-ranked faculty may, on average, be better teachers and thus deserve higher ratings. Teachers may also be less effective in large than small classes and, consequently, receive lower ratings (i.e., not necessarily because students take out their disdain for large classes by assigning lower ratings). In addition, students in upper-level courses in their major are most likely more motivated to take the course than students in lower-level general education courses.

If institutions are concerned that class size, student and course level, and student motivation to take a course could affect student ratings, we recommend they consider two options: They could create comparison groups of courses within the institution that control for these factors or consider adopting student ratings systems that statistically control for these extraneous influences.

Instructor Variables Weakly Related to Student Ratings

Generally, the following variables tend to show *little or no* relationship to student ratings and in our judgment do *not* require control.

Age and Teaching Experience

In general, instructor age and years of teaching experience are not correlated with student ratings. However, where weak correlations have been found, they tend to be negative (i.e., older faculty receive somewhat lower ratings; Feldman 1983; McPherson and Jewell 2007; Renaud and Murray 1996). Marsh and Hocevar (1991) pointed out that most of the studies of these variables have been cross-sectional comparisons of faculty cohorts that represent different age groups. In a *longitudinal* study, Marsh and Hocevar (1991) analyzed ratings of the *same* instructors across 13 years and found *no* systematic changes within instructors over time.

Centra (2009) found that first-year teachers tend to receive lower ratings than experienced assistant professors and higher-ranked faculty. He concluded that the lower ratings do not point to bias but probably reflect differences in experience and teaching skills, because first-year faculty are most likely still learning how to teach. Such lack of experience may help to explain why the correlation between ratings and student achievement is higher for experienced full-time instructors than it is for those who have taught less than 1 year (Sullivan and Skanes 1974).

Gender of the Instructor

In a review of 14 laboratory or experimental studies (where students rated descriptions of *fictitious* teachers who varied in gender), Feldman (1992) found few gender differences in global ratings. However, in a few studies, male teachers received higher ratings. In a second review of 28 studies of global ratings—*involving actual student ratings of real teachers*—Feldman (1993) found a very weak average correlation between instructor gender and student ratings ($r = .02$) that favored female instructors. Women also received slightly higher ratings on sensitivity and on concern with student level of preparedness and progress ($r = .12$). However, ratings of male and female teachers did not differ meaningfully on other dimensions of teaching.

More recently, Schulze and Tomal (2006) surveyed 2,042 students from 117 departments in 77 liberal arts colleges using a 12-item survey. Four items dealt with student perception of male and female students and male and female professors. Both male (42.7 %) and female (57.3 %) students perceived male and female students to generally be treated as equally competent and their questions and opinions to generally be given equal value. Female professors were treated with equal respect as male professors, and they were viewed as equally competent.

Some researchers have reported a student-gender-by-instructor-gender interaction. Basow (2000) surveyed a limited number of mostly white students—47 male and 61 female—at a small, liberal arts college. Students were asked to describe their "best" and "worst" professor and to rate them on the Bem Sex-Role Inventory. For "best" professor, female students chose female professors more frequently than did male students. There were no student gender differences for choice of "worst" professor. On the Bem Inventory, students rated both the "best" male and female professors as androgynous: strong in both active-instrumental (masculine) traits and expressive-nurturing (feminine) traits.

Feldman (1993) found that female students tend to give higher ratings to female teachers and male students give higher ratings to male instructors. Centra (2009) found that female instructors received slightly higher ratings, especially by female students, but that these were *not* accompanied by higher student self-ratings of learning. He, nonetheless, concluded that gender was not a bias, because the higher ratings might have reflected differences in effectiveness of teaching style: Female instructors were more likely to employ discussion than lecture, and they appeared to be more nurturing to their students (i.e., possibly more student centered). Regardless, the effect due to gender, although statistically significant, was so small that it would most likely not affect personnel decisions (see also Centra and Gaubatz 2000).

Race of the Instructor

Centra (1993) and Huston (2005) found, as we did, few studies of student ratings and instructor race conducted in North America. Centra (1993) speculated that students of the same race as the instructor *might* rate the instructor higher. However, in a doctoral dissertation using IDEA student ratings, Li (1993) found *no* differences between Asian and American students in their global ratings of (presumably Caucasian) instructors. Beyond the need for additional research, administrators and faculty who suspect racial or gender bias should incorporate additional indicators of teaching effectiveness into their teaching evaluations. Many such indicators are described in a later section of this chapter.

Personal Characteristics

Few personality traits have been found to correlate with student ratings (Braskamp and Ory 1994; Centra 1993). Using instructor self-report (e.g., personality inventories,

self-description questionnaires) as a criterion measure, Feldman (1986) found that only two (out of 14) traits had average correlations with a global teaching item that approached practical significance: positive self-esteem (r=.30) and energy/enthusiasm (r=.27). Murray et al. (1990) found significantly different patterns of correlations between personality traits and student ratings among psychology instructors teaching six different types of courses (e.g., introductory, graduate). The authors concluded that instructor personalities tended to be differentially suited to different types of courses. In a follow-up study using the same measures of personality, Renaud and Murray (1996) found positive correlations between average scores on a 10-item student ratings scale and colleagues' ratings of the instructor's orderliness (.65), defined as being neat and organized and disliking clutter and confusion. Working for the approval and recognition of others was also positively correlated (.56) with teaching effectiveness.

The few personality traits (i.e., positive self-esteem, energy/enthusiasm, orderliness, working for approval/recognition) that are related to student ratings tend to also enhance teaching effectiveness, and we therefore suggest they should *not* be controlled. One might expect, for example, that displaying energy/enthusiasm could stimulate student interest and that adopting orderliness could enhance classroom structure, which are teaching styles associated with student learning (Hoyt and Lee 2002a). Others disagree, arguing that instructor personality traits should *not* be correlated with student ratings, and if they are, then it indicates a bias in student ratings (e.g., Clayson and Sheffet 2006; Feeley 2002; Jenkins and Downs 2001; Patrick 2011).

What matters more than personality, however, is how the instructor's personal characteristics are manifested in the classroom. Most of the relationship between instructor personality and student ratings can be explained by the behaviors the instructor exhibits *when teaching* (Erdle et al. 1985). Put simply, the effect of instructors' personalities on ratings "may be caused more by what they do in their teaching than by who they are" (Braskamp and Ory 1994, p. 180).

Instructor First Impression

The first impressions an instructor makes in a course are important (Dorn 1987). In fact, impressions made within the first 2 weeks of the class are more important in determining end-of-course student ratings than is the instructor's general reputation (Buchert et al. 2008). Specifically, in separate studies, Buchert et al. (2008) compared end-of-course student ratings with ratings taken the second week of class (Study 1) and the first day of class (Study 2). In the first study, no significant differences were found between ratings collected during the second week of class and at the end of the course on items concerning the instructor's interest in the course, communication of the importance of the subject matter, expectations for student achievement and behavior, and grading criteria. The implication is that effective instructors do these things within the first 2 weeks of class. In Study 2, end-of-course ratings were higher than those taken during the first day of class on 14 of 18 items, including overall ratings of the course and instructor and ratings of how much students learned. So, impressions made the first day did not necessarily determine ratings at the end of the course.

Nonetheless, Dorn (1987) recommends that instructors systematically prepare for the first day of class and clearly communicate the purpose and content of the course, course objectives, course syllabus, assignments, readings, examinations, class rules (including policy against cheating and plagiarism), relevant background information about themselves, and expectations for student achievement. Instructors should also allow time for questions and for students to get to know each other. Finally, some content should be taught on the first day. As Svinicki and McKeachie (2011) advise, "An important function of the first day's meeting in any class is to provide…structure, that is, to present the classroom situation clearly, so that the students will know from the date of this meeting what you are like and what you expect" (p. 21).

Research Productivity

Research productivity has only a weak correlation with student ratings (Centra 1993). In his review, Feldman (1987) found an average correlation of .12 between research productivity and ratings of overall teaching effectiveness. Marsh and Hattie (2002) reported similar results ($r = .03$) in a survey of faculty across multiple departments in a single institution across a 3-year period. The near-zero correlation between teaching and research did not differ meaningfully from one department to another. These very low correlations suggest that being a productive researcher (i.e., devoting more time to research) is indicative of *neither* good *nor* bad teaching.

Student Variables Unrelated to Student Ratings

Researchers have examined possible bias in the following student variables and have found them unrelated to student ratings.

Age of the Student

Student age has little correlation with ratings (McKeachie 1979; Centra 1993). Of greater relevance is having experience as a college student, which brings more accomplished work habits, a characteristic that is positively correlated with ratings (see subsequent section on "Student Motivation").

Gender of the Student

Feldman (1977, 1993, 2007) reported no consistent gender effect, although some have reported a student-gender-by-instructor-gender interaction (see earlier section on instructor variables). In a comprehensive study of gender, Centra and Gaubatz

(2000) analyzed actual student ratings (rather than data from simulations) across a large number of 2- and 4-year institutions, involving a variety of academic disciplines. They found some gender preferences, particularly female students for female instructors. Although the differences were statistically significant, they were not large and would most likely not impact personnel decisions. Centra and Gaubatz (2000) speculated that the higher ratings female instructors received from female students, and sometimes from male students, might have reflected student preferences for certain teaching styles. Women in their study were more likely than men to use discussion than lecture, and they were perceived as more nurturing to students, as reflected in their scores on certain rating scales.

Level of the Student

Student level (e.g., first year, senior) has little practical effect on ratings (Davis 2009; McKeachie 1979). However, with experience, students gain better work habits, and they most likely become more interested in a content area as they move into upper-level specialized courses in their major. This can, in turn, affect their motivation to take the course and their effort in a class, factors that are positively correlated with student ratings (Hoyt and Lee 2002a) (see subsequent section).

Student Grade-Point Average

Davis (2009) concluded there is little or no relationship between student ratings and grade-point average, citing the empirical investigations of several authors (Abrami 2001; Braskamp and Ory 1994; Centra 1993; Marsh and Dunkin 1997; Marsh and Roche 2000; McKeachie 1997).

Student Personality

No meaningful relationships have been found between student personality and ratings (Abrami et al. 1982b). In our search of the literature, we found no relevant articles published in psychology journals on the topic of student ratings and student personality.

Course and Administrative Variables Unrelated to Student Ratings

Researchers have examined possible bias in the following course and administrative variables and have found them unrelated to student ratings.

Time of Day

The limited amount of research on the topic indicates the time of day the course is taught has no meaningful influence on student ratings (Aleamoni 1981; Feldman 1978).

Time During the Term When Ratings Are Collected

Any time during the second half of the term seems to yield similar results (Feldman 1979). Costin (1968) found no difference in ratings administered at the end versus the middle of the semester; others (Carrier et al. 1974) found no difference between ratings administered the last week versus the day of the final examination (although we recommend against administering ratings at a time when students are most concerned about an exam). Finally, Frey (1976) found no difference in ratings administered the last week of class versus the first week of the next semester. For ratings administered online, the rating period should end before the final exam is administered so that students' anxiety about or performance on the exam does not influence their responses (Hativa 2013a).

The research cited thus far suggests that many variables suspected of biasing student ratings are *not* correlated with them to any practically significant degree. However, research suggests the following variables are correlated with student ratings and *may require control*.

Instructor Variables Related to Student Ratings That May Require Control

Faculty Rank

Regular, full-time faculty members tend to receive higher ratings than graduate teaching assistants (Braskamp and Ory 1994). This variable may *not* require control because regular faculty as a group should be more experienced and, therefore, more effective teachers. In that sense, the higher ratings they receive probably reveal such experience and effectiveness.

Expressiveness

The Dr. Fox effect—where a professional actor, who delivered a dramatic lecture but with little meaningful content, received high ratings—suggested that student ratings might be influenced more by an instructor's style of presentation than by the substance of the content (Naftulin et al. 1973; Ware and Williams 1975; Williams and

Ware 1976, 1977). Naftulin et al. (1973) first reported the Dr. Fox effect, also known as "educational seduction," on an audience of educators and mental health professionals who viewed a videotaped lecture that expressively conveyed deliberately nonsensical information. Ware and Williams (1975) attempted to extend these findings by conducting a similar experiment with undergraduate and graduate students. They manipulated instructor expressiveness by having the actor either be enthusiastic, friendly, and charismatic (high expressiveness) or humorless and unenthusiastic (low expressiveness). They manipulated content (high, medium, low) by modifying the number of unrelated examples, circularity, and meaningless information contained in the lecture. Students completed ratings and a quiz following the lecture. Ware and Williams found that instructor expressiveness interacted with lecture content to influence student ratings. When instructor expressiveness was low, high lecture content led to higher ratings than low content. When expressiveness was high, lecture content did not affect ratings, but it did affect student achievement.

The literature generated by the Dr. Fox study was complex (see Perry et al. 1979a, b; Abrami et al. 1982a). Perry et al. (1979a) extended Williams and Ware's (1975) findings by doing several things to contrive a more typical classroom setting. They produced videotaped lectures performed by an actual college professor rather than a professional actor; they manipulated student incentive by allowing or disallowing students to earn credit for good quiz performance; they varied whether or not students had the opportunity to study lecture materials. Perry et al. (1979a) concluded that educational seduction differentially affected ratings and achievement only with high-incentive students (those told they could earn credits if they did well on the later quiz), but not with low-incentive students. When instructor expressiveness is low, content affects both ratings and achievement; when expressiveness is high, content affects only achievement. Perry et al. (1979a) concluded that educational seduction might be less likely to occur in conditions that more closely resemble typical college classes. In a second study, Perry et al. (1979b) found that instructor reputation interacts with expressiveness but not content. Students rated positive reputation, high-expressive professors more favorably than negative reputation, high-expressive professors.

In conducting a meta-analysis of 12 independent studies of educational seduction research, Abrami et al. (1982a) concluded that "(a) some teacher characteristics affect ratings and achievement similarly and (b) some teacher characteristics affect ratings and achievement differently or not at all" (p. 459). More to the point, Marsh and Ware (1982) concluded that manipulations of instructor expressiveness primarily influence ratings of instructor enthusiasm; manipulations of lecture content influence ratings of instructor knowledge, as well as student exam performance. Moreover, when student extrinsic motivation to achieve in a course is low, the influence of instructor expressiveness is substantial. Being more expressive produces higher student ratings *and* higher examination performance. In short, making the class interesting as well as informative helps students pay attention, especially when they are less motivated. Expressiveness, therefore, tends to enhance learning, and—using Marsh's (2007) definition of bias—we suggest it does *NOT* require control.

The fact that expressiveness has a moderate affect on student ratings should not be surprising given what is known about relationships between nonverbal behaviors and speaker credibility and persuasiveness (Burgoon et al. 1990). Such behaviors as vocal pleasantness (fluency and pitch variety), facial pleasantness, and facial expressiveness are positively associated with greater perceived competence. Greater perceived persuasiveness is linked to greater vocal pleasantness, facial expressiveness, and kinesthetic relaxation. In light of this, we agree with Abrami et al.'s (1982a) recommendation that administrators should use ratings to make only crude judgments of teaching effectiveness and that ratings should be collected from multiple classes and used in combination with other data when making promotion and tenure decisions.

Student Variables Related to Student Ratings That May Require Control

Student Motivation

Instructors are more likely to receive higher ratings in classes where students had a prior interest in the subject matter (Marsh and Dunkin 1997) or were taking the course as an elective (Aleamoni 1981; Braskamp and Ory 1994; Centra 1993; Feldman 1978). Although Marsh (2007) concluded the reason for taking a course (which overlaps with student motivation) is related to student ratings, this variable is *not a bias* because motivated students are likely to learn more. However, because motivation to take the course is a student characteristic and *not* necessarily a reflection of the instructor's teaching effectiveness, we believe this variable requires some control. Control could be established by creating comparison groups of courses within the institution that control for student self-reported motivation or by statistically adjusting ratings for student motivation.

Possibly related to this, Centra (2009) found that required courses tend to receive lower ratings than other kinds of courses, but the differences are not great. Nonetheless, it would *not* be fair to penalize instructors who teach required courses or appropriate to reward those teaching an elective course. Expressing another perspective, Hoyt and Cashin (1977) found that some "required" courses are very popular with students (especially required courses in the major) and some "elective" courses are regarded less positively (especially science or mathematics electives taken to satisfy distribution requirements). Measures of student motivation/interest in the course have therefore been shown to be more accurate as a control variable.

Expected Grade

In a study involving over 50,000 classes, Centra (2003) examined the relationship between the grade students expected to receive in a course and ratings of the

quality of instruction. Controlling for class size, teaching method, and student ratings of progress on learning outcomes, expected grade generally had no effect on ratings across eight subject-matter areas. However, others have reported positive but low correlations (.10 to .30) (Braskamp and Ory 1994; Centra 2003; Feldman 1976a; Howard and Maxwell 1980, 1982; Marsh and Dunkin 1997; Marsh and Roche 2000).

Three possible hypotheses have been proposed for these low positive correlations. The *validity hypothesis* posits that students who learn more earn higher grades *and* assign higher ratings (which supports the validity of student ratings). The *leniency hypothesis* asserts that instructors who give higher grades than the students deserve receive higher ratings than the instructor deserves. A third hypothesis is that *student characteristics* (e.g., high interest or motivation) lead to greater learning and, therefore, higher grades *and* higher ratings.

In two studies of IDEA data, Howard and Maxwell (1980, 1982) concluded that students' self-reported learning and desire to take the course explained most of the shared variance between expected grades and global ratings of the instructor, which supports the validity and student characteristics hypotheses. More recently, Marsh (2007) reviewed studies that supported the validity hypothesis, with some support for student characteristics.

Additional support for the validity hypothesis comes from Baird (1987) who argued that the imperfect and inconsistent correlations between student ratings and course grades stem from discrepancies between what a student actually learned and the grade received. Baird (1987) found that students' perceptions of how much they learned in the course were highly correlated with overall ratings of the professor and course, but anticipated letter grade was only weakly correlated. Moreover, the relationship between anticipated letter grade and actual course grade was especially weak ($r = .18$). The author concluded, "These results support the validity hypothesis regarding the grades-rating correlation. Students rate instructors according to how much they believe they have learned rather than according to their anticipated grades" (p. 91).

McKeachie's (1979) position on this matter still seems appropriate: "[I]n courses in which students learn more the grades should be higher and the ratings should be higher so that a correlation between average grades and ratings is not necessarily a sign of invalidity" (p. 391). To control for the possibility of grading leniency, however, one might have peers (faculty knowledgeable in the subject matter) review the instructor's course material, especially exams, test results, graded samples of essays, projects, grade distributions, and so forth to judge the course standards and the bases for grading in the course (McKeachie 1979).

Ultimately, if an instructor gives higher grades than students deserve to receive higher ratings than the instructor deserves, we question whether the problem lies with student ratings or with the instructor. As Hativa (2013b) points out, "one cannot blame SRIs if the real issue/problem is unethical teacher behavior" (p. 60). Moreover, the faculty member who wants to increase course ratings would be better served by practicing other more productive behaviors than assigning lenient grades (Hativa 2013b). Teaching effectively—by challenging students and stimulating

their interests (Marsh and Roche 2000)—and responding proactively to student feedback about instruction and the course (Centra 2003) are more likely than leniency to lead to greater student learning and higher ratings.

Course Variables Related to Student Ratings That May Require Control

Class Size

Although there is a tendency for smaller classes to receive higher ratings, it is a very weak inverse relationship (average $r = -.09$) (Feldman 1984). Hoyt and Lee (2002a) found that the effect of class size on ratings was not always statistically significant, but when it was, the relationship was negative. Instructors teaching small classes therefore have a slight advantage over those teaching large classes.

There may be several reasons for the advantage of small over larger classes. Students may be more likely to interact with classmates, speak up in class, ask questions, and establish a relationship with the instructor. Instructors may require lengthier writing assignments, more graded homework, and more essay exams, all of which could lead to greater student learning (Hativa 2013b; Marsh 1987). Under these conditions, if students perceive that they learn more, they most likely assign higher ratings.

The effect of class size is most apparent when comparing very large with small classes. Benton and Pallett (2013) found that in very large classes, instructors were more likely to emphasize factual knowledge and less likely to develop communication skills. In turn, students were less likely to report progress on communication skills and creative capacities, such as writing, inventing, designing, and performing. The type of learning where students in very large classes approached the progress of those in small and medium classes was in developing basic background in the subject matter.

Centra (2009) found that smaller classes not only tend to receive higher ratings, but that students in those classes report learning more. Because class size is related to both student learning and effective teaching, it is, therefore, *not* considered a bias. However, Centra suggested that institutions might want to take size into consideration—by using comparative data—when considering student ratings in personnel decisions. Alternatively, adjusting statistically for class size, as is done in IDEA student ratings, is another possibility.

Level of the Course

Although we reported previously that *level* of the student is unrelated to student ratings, higher-level courses (especially graduate courses) are rated somewhat higher than lower-level courses (Aleamoni 1981; Braskamp and Ory 1994; Feldman 1978).

However, the differences tend to be small. Such differences can most likely be explained by greater student motivation and work habits in graduate courses. However, statistical controlling for level of the course would most likely result in substantial grouping error, as not all lower-level students lack motivation/work habits and not all upper-level students are high on those characteristics. Adjusting for individual student self-reported motivation/work habits is a preferred way to reduce possible bias.

Locally, institutions should check to see if lower-level classes receive lower ratings than upper-level classes. Similarly, they should compare undergraduate with graduate classes. If differences exist, do they remain after controlling for student motivation/work habits and class size? If so, we recommend developing local comparative data for the appropriate levels.

Academic Discipline

Feldman (1978) reviewed studies showing that courses in the humanities and arts receive higher ratings than social sciences, which in turn receive higher ratings than math and science. Others (Braskamp and Ory 1994; Cashin 1990; Centra 1993, 2009; Hoyt and Lee 2002b; Kember and Leung 2011; Marsh and Dunkin 1997; Sixbury and Cashin 1995) found similar results. Although there is increasing evidence that ratings differ between disciplines, it is *not clear* why.

Cashin (1990) suggested some possible explanations. For example, some fields may be rated lower because they are more poorly taught; if so, then these differences do *not* require control. There is some evidence, for example, that mathematical/science courses tend to receive lower ratings (Centra 2009; Hoyt and Lee 2002b). Ratings in those courses are lower for the teaching styles of stimulating interest, fostering collaboration, and encouraging student involvement, which are related to overall measures of teaching excellence. Notably, students also rate the courses in these content areas more difficult than other courses, and they express less motivation to take the course (Hoyt and Lee 2002b).

If instructors in fields requiring more quantitative reasoning skills are rated lower because today's students are less competent in such skills—another hypothesis offered to explain why some disciplines are rated lower (Cashin 1990)—then some control is necessary. Centra (2009) suggested that institutions might want to use comparative data to determine if the lower ratings result from lower student quantitative skills.

Another explanation for disciplinary differences is the sequential/hierarchical structure of content in some disciplines (Hativa 2013b). *Hard* disciplines (see Biglan 1973), for example, have a structured knowledge sequence organized around a theory agreeable to all members of the field (e.g., engineering, chemistry). Consequently, students must have a solid knowledge base in prior courses to succeed in subsequent ones. There may also be differences in the kinds of faculty and students attracted to certain disciplines. Some faculty in the hard sciences, for example, may be more attracted to a certain discipline because of research

interests than for teaching opportunities. In addition, some students in specific fields might possess common attitudes toward and expectations about how courses should be taught.

Another possible explanation for disciplinary differences is the type of teaching methods employed in the classroom (Hativa 2013b). Instructors in soft disciplines, for example, tend to exhibit a wider range of teaching behaviors than those in hard disciplines (Franklin and Theall 1992). The authors found that instructors in the arts and humanities more frequently set objectives at the mid and upper levels of Bloom's taxonomy of cognitive objectives and used active teaching methods, whereas those in science, technology, engineering, and mathematics (STEM) courses rely more on lower-level objectives and employ lecture more predominately. In addition, instructors in STEM fields employ teaching methods associated with student learning less frequently than do those in non-STEM courses (Benton et al. 2012b).

In spite of some differences, we agree with Hativa (2013a) that unique forms should not be created for different disciplines. Student perceptions of teaching behaviors and their relationships with learning outcomes are similar across disciplines (Marsh and Dunkin 1997; Murray 2007). Moreover, consistency has been found across disciplines in what constitutes an effective teaching and learning environment (Kember and Leung 2011).

Workload/Difficulty

Some instructors fear ratings may be biased because students perceive some disciplines as more difficult than others. For example, students tend to rate natural science courses the most difficult (Centra 2003; Hoyt and Lee 2002a). However, course workload and subject-matter difficulty are only weakly correlated with student ratings (Centra 1993, 2003; Marsh 2001; Marsh and Roche 2000). Contrary to what some might expect, the correlations are positive—students give somewhat *higher* ratings to difficult courses that require hard work. Still, the correlations are not large. Greenwald and Gillmore (1997) reported just the opposite—that courses with lighter workloads received *higher* student ratings. However, Marsh (2001) reanalyzed their data and found two nearly *un*correlated components of workload: "bad workload" (time spent that was *not* valuable) and "good workload" (i.e., time spent on activities related to instructional objectives). Whereas "bad workload" was correlated negatively with student ratings, "good workload" (work that helps students learn) was positively correlated.

The effect of subject-matter difficulty on student ratings may depend on the type of learning emphasized in the course. Hoyt and Lee (2002a) controlled for the instructor's influence on student perceptions of the difficulty of the subject matter. They computed a residual score that represented the students' perception of difficulty once the instructor's influence (e.g., amount of reading and non-reading assignments students reported) had been removed. If students perceived the discipline as difficult, ratings were usually slightly lower. However, difficulty was *positively* correlated with student progress on basic cognitive objectives related to factual knowledge and learning of principles and theories.

The key for instructors is to find the right amount of difficulty. A few researchers (Centra 2003; Marsh and Roche 2000; Marsh 2001) have reported a nonlinear relationship between workload/difficulty and student ratings. For example, Centra (2003), using a large database of classes, found that courses were rated lower when they were perceived as either too difficult or too elementary; the highest ratings were found in classes where difficulty/workload was rated as "just right." However, the relationship was not strong.

To sum up this section, relatively few variables are related to student ratings that are not also correlated with instructional effectiveness. Nonetheless, a few student and course variables may require some control. In the following paragraphs, we address administration procedures that can affect student ratings when not controlled.

Manipulating Administrative Procedures

The validity of student ratings depends not only on the quality of the instrument but also on how properly the ratings are administered (Beran et al. 2007; Hativa 2013b). Several administrative factors may affect ratings, as described in the following paragraphs.

Non-anonymous Ratings

Students tend to give higher course and instructor ratings when they surrender their anonymity by signing the ratings (Braskamp and Ory 1994; Centra 1993; Feldman 1979; Marsh and Dunkin 1997). Requiring students to sign their names may inflate the ratings because some students may be concerned about possible reprisals.

With the growing trend in learning analytics (Brown 2012; Dyckhoff 2011), institutions need to conduct analyses on multiple sources of student data. In some cases, it may be desirable to correlate ratings with other indicants of student learning and development. This would especially be true with midcourse ratings taken during specific class periods to examine correlations between ratings of teaching methods and quizzes or other learning outcomes. In such situations, ratings might need to be kept confidential but not anonymous. Nonetheless, instructors should urge students not to sign their ratings. They should assure students that their responses are confidential and that only aggregated data and typed comments will be presented to the instructor, *and only after grades have been submitted*.

Instructor Present While Students Complete Ratings

Ratings tend to be higher (Braskamp and Ory 1994; Centra 1993; Feldman 1979; Marsh and Dunkin 1997) when the instructor is present, possibly for the same

reason as non-anonymous ratings. We recommend that the instructor leave the room, and a neutral person collect the ratings.

Purpose of the Ratings

Some researchers have investigated whether the perceived purpose of conducting ratings affects students' responses. Centra (1976) found ratings of the instructor's overall effectiveness did not differ between administrative conditions that specified ratings would be used for *personnel decisions* versus those that said they would be used only for improvement. In reviewing Centra's (1976) results, however, Feldman (1979) noted that the effect of instructions on ratings varied by the teacher. In some cases, specifying that the ratings would be used for tenure, salary, and promotion decisions resulted in higher ratings, whereas in others it had no effect or was associated with lower ratings. So, the effect of varying the directions on student ratings is small (Marsh 2007) and inconsistent. We suggest that instructors include in the standard directions the intended purpose(s) of the ratings. Although this will *not* eliminate potential bias, it will control *variations* in ratings due to differences in student beliefs about how they will be used.

Analyzing the Underlying Dimensions of Ratings

There is broad agreement that student ratings are multidimensional (i.e., that they reflect several different aspects of teaching). The number of dimensions varies depending, in part, on the form studied and the number and kind of individual items it contains. Put simply, multidimensionality suggests *no single student ratings item or set of related items is useful for all purposes*.

There have been a number of factor-analytic studies conducted (see Abrami and d'Apollonia 1990; Hoyt and Lee 2002a; Kulik and McKeachie 1975; Marsh and Dunkin 1997) in which the dimensions were derived statistically. Both Centra (1993) and Braskamp and Ory (1994) identified six factors commonly found in student ratings forms: course organization and planning, clarity and communication skills, teacher-student interaction and rapport, course difficulty and workload, grading and examinations, and student self-reported learning. Employing confirmatory factor analysis, Marks (2000) reported five dimensions: (1) organization, (2) workload/difficulty, (3) expected/fairness of grading, (4) liking/concern, and (5) perceived learning. Marsh's (1984, 2007) SEEQ has nine components: learning/value, enthusiasm, organization, group interaction, individual rapport, breadth of coverage, exams/grades, assignments, and workload. Other student ratings instruments have items measuring some or all of the above dimensions. Hoyt and Lee (2002a) reported five dimensions of teaching based on the *IDEA Diagnostic Form*: (1) providing a clear classroom structure, (2) stimulating student interest, (3) stimulating student effort, (4) involving students, and (5) student interaction.

In several of his reviews of the literature, Feldman (1976b, 1983, 1984, 1987, 1988) categorized student ratings items (and gave examples) into as many as 22 different logical dimensions. In a later review, Feldman (1989b, 2007) identified 28 dimensions.

In her review of numerous studies, Hativa (2013a) proposed a three-level model of the dimensions underlying effective teaching. At the top level is general instructional skill, which is supported by the view that student ratings measure a single overall component (d'Apollonia and Abrami 1997). At the next level, two orthogonal dimensions distinguish between cognitive (i.e., communicating content) and affective (i.e., interpersonal) aspects of teaching. Hativa (2013a) reported that a number of authors have found support for these two dimensions (Addison and Stowell 2012; d'Apollonia and Abrami 1997; Hativa et al. 2010). Finally, at the subordinate level, three teaching behaviors comprise the cognitive dimension—lesson clarity, course and lesson organization, and engaging/interesting presentations. Within the affective dimension are interactions/questioning/answering and rapport with students. Other studies support the validity of these low-level dimensions (Braskamp and Ory 1994; Marsh 2007).

The consistent multidimensionality found in ratings suggests students can distinguish among factors related to teaching effectiveness. Moreover, students can differentially weight teaching behaviors when making overall evaluations of the instructor. When using student ratings data to improve teaching, instructors should distinguish among the various items and their factor structure to insure that all of the appropriate dimensions of teaching are rated. Hoyt and Lee (2002a) found that the relevance of 20 different IDEA teaching methods varied depending upon which learning objectives were emphasized in a course. The implication was that different kinds of learning require different types of teaching.

An anonymous reviewer of this chapter offered yet another recommendation that dimensionality be tied to the instructor's objectives for the course. Teaching effectiveness could then be defined as student ratings of progress on objectives the instructor identifies as important rather than by specific dimensions of teaching. The IDEA Student Ratings of Instrument System (http://www.theideacenter.org/) takes this approach. Instructors rate each of 12 learning objectives as either *essential*, *important*, or of *no or minor importance* to the course. Students are then asked to rate the amount of progress they made on each objective, ranging from 1 = *no apparent progress* to 5 = *exceptional progress*. Students report greater progress on objectives the instructor identifies as either essential or important. Moreover, class average progress ratings are highly correlated with overall ratings of the course and instructor (Hoyt and Lee 2002a).

Although there is general agreement that student ratings are multidimensional and that various dimensions should be used when their purpose is to improve teaching, there is disagreement about how many and which dimensions should be used for personnel decisions (Apodaca and Grad 2005; Harrison, Douglas, and Burdsal 2004; Hobson and Talbot 2001; Renaud and Murray 2005). In several articles, Abrami (e.g., Abrami and d'Apollonia 1991) suggested that one or a few global/summary items might be sufficient for personnel decisions. Others have

made a similar recommendation (e.g., Braskamp and Ory 1994; Cashin and Downey 1992; and Centra 1993). Harrison and colleagues also confirmed that various weighted and unweighted measures of overall evaluations of teaching effectiveness are highly intercorrelated (Harrison et al. 2004).

Offering another view, McKeachie (1997) argued that when it comes to personnel decisions, student ratings of progress on educational goals and objectives are preferable to multiple dimensions or a single measure of overall teaching effectiveness. Effective teaching can be demonstrated in many ways, and no instructor should be expected to demonstrate proficiency in all methods and styles. Moreover, teaching methods may vary, depending upon the course content, student characteristics, and size of class. In other words, no single set of teaching behaviors "constitute[s] a necessary-and-sufficient condition" for effective teaching (Hativa 2013a, p. 33). Regardless of which measures are used, administrators and members of personnel committees should use broad categories (e.g., exceeds expectations, meets expectations, fails to meet expectations) rather than try to interpret decimal point differences (d'Apollonia and Abrami 1997; McKeachie 1997; Pallett 2006).

The research cited thus far has summarized evidence of the validity of student ratings as found in correlations with student achievement, correlations with other criteria, examinations of potential bias, manipulations of administration procedures, and factor-analytic studies. In the next sections, we summarize research comparing ratings administered online versus on paper and in online versus face-to-face classes.

Student Ratings Administered via Paper and Pencil Versus Online

Web-based student ratings of instruction are increasing due to their efficiency and lower administration costs when compared to paper-and-pencil surveys (Avery et al. 2006; Dommeyer et al. 2004). Online formats may be less susceptible to faculty influence (Anderson et al. 2005; Dommeyer et al. 2004) and are consistent with campus sustainability goals. They offer more flexibility in making modifications to questions or survey design, can be easily integrated into an online course management system, and can be completed via mobile technology at times convenient to students (Hativa 2013a). Moreover, online directions and procedures can be uniform for all classes, further enabling instructors to be less involved in the administration process (Layne, DeCristoforo, and McGinty 1999).

Online delivery offers other advantages over paper-and-pencil administration. Because students can respond outside of class at their convenience, it frees up class time for other activities (Dommeyer et al. 2004; Layne et al. 1999). Response rates to open-ended questions posted online tend to be higher (Johnson 2003), and written comments are typically lengthier (Hardy 2003; Johnson 2003; Layne et al. 1999) and more detailed (Alhija and Fresko 2009). Possible explanations for the better quality of online comments include lack of time constraints, student preference for keyboarding or texting over handwriting, and greater trust in confidentiality (Hativa 2013a).

The downside is that student response rates are typically lower for online formats (e.g., Avery et al. 2006; Layne et al. 1999), although lower response rates do not necessarily result in lower mean ratings (Avery et al. 2006; Benton et al. 2010a; Layne et al. 1999). Leung and Kember (2005) found that after controlling for relevant demographic variables, students who could choose either paper-and-pencil or online formats did not differ in their ratings.

Lower response rates occur for several reasons, among them student concern about anonymity, computer technical difficulties, and the time required to respond outside of class (Dommeyer et al. 2004). Students who earn a low grade or no grade are less likely than others to respond; in contrast, students in their major are more likely to do so (Adams and Umbach 2012). Some instructors may fear lower response rates that create a negative bias because students who are dissatisfied with the course or instructor might be more likely than others to complete the ratings (Johnson 2003). However, correlations between response rate and overall ratings of the instructor and course are, on average, quite low (Benton et al. 2010a; Johnson 2003), which suggests response bias is less likely.

In spite of the disparity in response rates, researchers have consistently found no meaningful differences in student ratings delivered online versus on paper. The studies reviewed here included a mixture of students enrolled in courses conducted on campus and online. When the same students respond both online and on paper, the correlations between their global ratings of the instructor (.84) and course (.86) are high (Johnson 2003). Further, no meaningful differences are found in individual item means, number of positive and negative written comments (Venette et al. 2010), scale means and reliabilities, and the underlying factor structure of the ratings (Leung and Kember 2005). Similarly, when different students respond to online and paper surveys, no meaningful differences are found in student progress on relevant course objectives, global ratings of the course and instructor, frequency of various teaching methods (Benton et al. 2010a), subscale means (Layne et al. 1999), the proportion of positive and negative written comments (Hardy 2003), and the underlying factor structure (Layne et al. 1999).

Suggestions for Increasing Online Response Rates

Higher online response rates are more likely when instructors clearly communicate their expectations for compliance and when students complete ratings for more than one course (Johnson 2003). Other recommendations include ensuring student confidentiality, monitoring response rates, encouraging instructor follow-up, sending reminders, acknowledging and rewarding instructors with high response rates, and integrating the process into the campus culture (see The IDEA Center 2008). In addition, Linse (2012) recommends mentioning improvements made to the course based on feedback from previous SRIs, guiding students in how to write helpful comments, building rapport with students, creating

a culture of assessment by collecting other types of feedback during the course, reserving a room where students can complete ratings online, and making it clear feedback is valued. With the availability of mobile technology, tablets, and laptop computers, students can complete online ratings during a class period, which should also enhance response rate.

Student Ratings in Face-to-Face Versus Online Courses

The online classroom environment has characteristics that may either diminish or enhance student learning. Opportunities for student participation may either be reduced or increased, depending on how the course is structured; access to the instructor may decline or improve, depending upon instructor responsiveness to e-mail and student posts; students may either moderate or expand connections with classmates, depending on how frequently they post comments or participate in "chat rooms." All of these elements could affect student ratings either positively or negatively (Smith et al. 2000). Because of differences between online and face-to-face classroom environments, some have investigated whether instructors can use the same student ratings instrument in both settings (e.g., Beattie et al. 2002; Benton et al. 2010b). The general finding is that student ratings collected in face-to-face and online courses are actually more similar than they are different.

One means for comparing ratings across class settings is to sample only students who complete an online student ratings form. Taking this approach, Benton et al. (2010b) found that student progress on relevant objectives, global ratings of the course, and the instructor and the frequency of various teaching methods were comparable between courses identified exclusively as either face-to-face or online (Benton et al. 2010b). When ratings collected on paper from students enrolled on campus are compared with ratings collected online from students enrolled in distance courses, individual item means, internal consistency reliabilities, and the underlying factor structures are very similar (McGhee and Lowell 2003). Furthermore, item means and the overall assessment of the instructor are nearly identical between students enrolled in multiple online and face-to-face sections of the same course taught by the same instructor (Wang and Newlin 2000).

Nonetheless, some differences do exist. As one might expect, response rates to online forms are somewhat lower in online than face-to-face courses. However, the correlations between response rate and overall ratings of the instructor and course are, on average, low, making negative response bias less likely (Benton et al. 2010b). Not surprisingly, students in online courses report greater instructor use of educational technology to promote learning, and such use is more highly correlated with student progress in online courses. In addition, students report somewhat more reading in online courses (Benton et al. 2010b).

Usefulness of Student Ratings

Poor practice has perhaps led to greater invalid and inappropriate use of ratings than has all purported sources of bias combined (Hativa 2013a). Faculty misuse ratings when they divert from standard administration procedures, grade leniently based on the erroneous belief it will increase ratings, make course decisions about objectives and content solely on the basis of student feedback, fail to respond at all to student feedback, and disparage ratings as having no value whatsoever. Administrators misuse student ratings when they make them the sole basis for evaluating teaching effectiveness, make decisions on the basis of a single class, overemphasize small differences in ratings, ignore the role of extraneous factors, fail to consider comparative data, rely too heavily on student written comments, and fail to respond to student feedback (Hativa 2013a).

Sometimes faculty are skeptical about whether student ratings feedback can have a positive impact on teaching improvement (Campbell and Bozeman 2007). Substantial evidence indicates that it can. In Cohen's (1980) meta-analysis of 17 studies, receiving feedback from ratings administered during the first half of the term was *positively* related to improving teaching as measured by student ratings administered at the end of the term. All classes in the meta-analysis had ratings administered during the first half of the semester and again at the end. Cohen used the end-of-term ratings as the measure of improvement. He found that student ratings feedback by itself modestly improved instruction. However, faculty made the most improvement in teaching when student feedback was combined with consultation.

Others have reported similar results (Brinko 1990; Hampton and Reiser 2004; Hativa 2013a; Knol 2013; Marincovich 1999; Marsh 2007; Marsh and Roche 1993; Ory and Ryan 2001; Penny and Coe 2004). For example, Knol (2013) employed a randomized block design to examine the effects of feedback only and feedback plus consultation on improving the quality of lectures at a Dutch University. Prior to teaching their course, professors were randomly assigned to either a feedback-only condition, where students provided feedback shortly after each of three lectures; a feedback-plus-consultation condition, in which student feedback was combined with consultation following each lecture; or a control condition in which student feedback came only at the end of the course. Knol found large effect sizes (Cohen's $d > .80$) for feedback plus consultation on faculty-reported gains in knowledge, focus on teaching, and plans for improvement. In addition, feedback with consultation had a strong impact on student ratings of the instructor's lecturing skills and students' ratings of how much they learned from the lectures.

Discussing ratings with a peer or consultant improves their usefulness (Aleamoni 1978; Marsh and Overall 1979), especially when such conversation targets problems identified by students (Marsh and Roche 1993). Faculty find especially helpful feedback about interaction with students, grading practices, global ratings of the course and instructor, and structural issues (e.g., pace of course, exam difficulty and content, and textbook) (Schmelkin et al. 1997).

In the absence of a consultant, instructors should reflect on what the ratings mean as a useful first step. Kember and colleagues developed a four-category scheme for assessing quality of self-reflection (Kember et al. 2008). In *nonreflection*, the instructor simply looks at the ratings without giving them much thought. At the second level of *understanding*, the instructor attempts to grasp what the ratings mean but does not relate them to his or her own experiences. It is not until *reflection* that instructors relate the results to their own experience teaching the specific course. Finally, in *critical reflection*, the teacher undergoes a transformation, perhaps brought on by the disequilibrium or cognitive dissonance produced when the feedback differs from the teacher's view of how things went.

Such feedback can be humbling, but it may lead instructors to admit that something in the course or their teaching needs to change (Weimer 2009). Meaningful change, according to instructors who have made significant improvements in end-of-course ratings, does not require great effort (McGowan and Graham 2009). Improvements in ratings are most frequently associated with creating opportunities for active learning in the classroom, fostering better student-teacher interactions, setting expectations and maintaining high standards, being prepared for class, and revising procedures for assessing student work (McGowan and Graham 2009).

Unfortunately, the actual use of student ratings for formative purposes falls far short of its potential. Pallett (2006) suggested three possible reasons. First, institutions sometimes place too much emphasis on the summative component of ratings. When student ratings are overemphasized for summative evaluation and underutilized for developmental purposes, faculty often lose trust in the process and see little or no benefit in collecting student feedback. Such misuse erodes the potential benefits of ratings and can create a negative climate for faculty evaluation. A second reason for underutilization is the challenge of creating valid and reliable ratings instruments that provide helpful feedback. Third, at some institutions, there is insufficient mentoring. Credible mentors who are trusted colleagues, not necessarily involved in personnel decisions, are needed to provide feedback and make recommendations for improvement.

Conclusion

There are probably more studies conducted on student ratings than on all other kinds of data used to evaluate college teaching combined. Although one can find individual studies that support almost any conclusion, for many variables, there are enough studies to discern trends. In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control, perhaps more so than any other data used for faculty evaluation. Moreover, they can help instructors improve their teaching, especially when combined with self-reflection and consultation.

Nonetheless, student ratings are *only one source* of data about teaching and must be used in combination with additional evidence if one wishes to make a judgment about all of the components of college teaching. Further, student ratings must be

interpreted. We should not confuse a source of data with the evaluators who use it—in combination with other kinds of information—to make judgments about an instructor's teaching effectiveness (Cashin 2003).

This chapter reveals that extensive research has been conducted on student ratings, including replication of key findings. Where should the field go from here? What questions remain unanswered or need to be asked? In the following paragraphs, we make a few suggestions.

Directions for Research and Practice

Perhaps the greatest need is educating administrators, faculty, and students about the value of student ratings data when used appropriately. Researchers can challenge misconceptions about ratings through articles in the press (e.g., *The Chronicle of Higher Education*, *Inside Higher Education*), e-mail (e.g., POD@listserv.nd.edu), blogs (e.g., IDEAblog, http://theideacenter.org/search/site/IDEAblog), Webinars, professional organizations (e.g., Professional and Organizational Development Network), and presentations at faculty or department chair development workshops. Because people do not easily discard misconceptions, researchers should investigate which strategies are most effective for bringing about conceptual change in instructor and administrator erroneous beliefs. Much research has been conducted in the field of cognitive psychology that could be applied to challenge misconceptions about student ratings.

Another exciting area of potential research is in the emerging field of learning analytics, defined as:

> an umbrella term for the use of data, statistical analysis, and explanatory and predictive models to gain insights and act on complex issues. As a genre of analytics, learning analytics (LA) uses these methods to achieve greater success specifically in student learning. LA can be used in a variety of ways, some of which include alerting faculty, students, and advisors when intervention is needed; providing input for continuous improvement in course design and delivery; and enabling personalization of the learning environment. (Brown 2012)

Institutions of higher education collect vast amounts of data on students and courses that could be integrated to enable better decision making. Researchers and faculty development experts should focus on how student ratings data might better inform teaching and learning. LA provides a way to monitor learner activity and progress and to then make predictions, which affect both student and teacher decision making. The outcome is that instructors can make "targeted improvements" midcourse, and students can receive "targeted alerts" on how they might improve. This model differs from the typical system where feedback does not come until the end of the course. LA enables such feedback while the course is still active.

Innovations in course design offer another possibility for future research. Among them are online courses, massive open online courses (MOOCS), and the flipped classroom. Online courses are becoming more prominent in higher education. The research reviewed in this chapter found no evidence of meaningful differences between student ratings administered online versus face-to-face. However, there may be teaching methods and strategies that online instructors

employ current student ratings instruments fail to assess. Moreover, some faculty create courses online they never teach or teach courses they do not develop (Creasman n.d.). What are the implications of this for faculty evaluation? In addition, researchers should continue to investigate methods for increasing student response rates to online surveys.

With the advent of MOOCS, class sizes can increase to tens of thousands of students. What affects do such large classes have on student ratings and student achievement? For example, student ratings of progress made in developing creative capacities and communication skills are about a standard deviation lower in classes of 100 compared to classes fewer than 15 (Benton and Pallett 2013).

The flipped classroom is yet another innovation in higher education. Students view lectures online in preparation for working on problems with other students in class. Researchers should examine whether typical student ratings instruments need to incorporate additional items to assess student learning and teaching methods in such blended learning situations.

Finally, more must be done to identify strategies for encouraging faculty and administrators to use students ratings effectively. How might faculty developers successfully engage faculty in reviewing student ratings feedback? At the institutional level, how might aggregated student ratings data be used to promote effective faculty development and continuous improvement, to augment program reviews, and to address accreditation standards?

So although research on student ratings seems exhaustive, many interesting questions remain unanswered. Whether the conclusions reached in this chapter hold true for all contexts is an empirical question. If an institution has reason to believe that a given conclusion does *not* apply, key players should gather local data and conduct research to address the issue. In the absence of evidence to the contrary, however, the following general conclusions can be used as a guide (Marsh 2007, p. 372):

> SETs [student evaluations of teaching effectiveness] are multidimensional, reliable and stable, primarily a function of the instructor who teaches a course rather than the course that is taught, relatively valid against a variety of indicators of effective teaching, relatively unaffected by a variety of potential biases, and are seen to be useful by faculty, students, and administrators.

Acknowledgment The authors wish to thank The IDEA Center for granting permission to use chapter portions of IDEA Paper No. 50, *Student Ratings of Teaching: A Summary of Research and Literature.*

References

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* (New directions for institutional research, no. 109, pp. 59–87). San Francisco: Jossey-Bass.

Abrami, P. C., & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (New directions for teaching and learning, no. 43, pp. 97–111). San Francisco: Jossey-Bass.

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness- generalizability of "N = 1"research: Comments on Marsh (1991). *Journal of Educational Psychology, 83*, 411–415.

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982a). Educational seduction. *Review of Educational Research, 52*, 446–464.

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

Abrami, P. C., d'Apollonia, S., & Rosenfeld, S. (2007). The dimensionality of student ratings of instruction: What we know, do not know, and need to do. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385–445). Dordrecht: Springer.

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.

Addison, W. E., & Stowell, J. R. (2012). Conducting research on student evaluations of teaching. In M. E. Kite (Ed.), *Effective evaluations of teaching: A guide for faculty and administrators* (pp. 5–12). Retrieved from the Society for the Teaching of Psychology website:http://teachpsych.org/ebooks/evals2012/index.php

Aleamoni, L. M. (1978). The usefulness of student evaluations in improving college teaching. *Instructional Science, 7*, 95–105.

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

Aleamoni, L. M. (1987). Student rating myths versus research facts. *Journal of Personnel Evaluation in Education, 1*, 111–119.

Alhija, F. N. A., & Fresko, B. (2009). Student evaluation of instruction: What can be learned from students' written comments? *Studies in Educational Evaluation, 35*, 37–44.

Anderson, H. M., Cain, J., & Bird, E. (2005). Online student course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education, 69*, 34–43.

Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Studies in Higher Education, 30*(6), 723–748.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*, 1198–1208.

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

Avery, R. J., Bryant, W. K., & Mathios, A. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations. *The Journal of Economic Education, 37*, 21–37.

Babor, T. F., & Del Boca, F. K. (1992). Just the facts: Enhancing measurements of alcohol consumption using self-report methods. In R. Litten & J. Allen (Eds.), *Measuring alcohol consumption: Psychosocial and biochemical methods* (pp. 3–19). Totowa: Humana Press.

Babor, T. F., Steinberg, K., Anton, R., & Del Boca, F. K. (2000). Talk is cheap: Measuring drinking outcomes in clinical trials. *Journal of Studies on Alcohol, 61*(1), 55–63.

Baird, J. S. (1987). Perceived learning in relation to student evaluation of university instruction. *Journal of Educational Psychology, 79*, 90–91.

Basow, S. A. (2000). Best and worst professors: Gender patterns in students' choices. *Sex Roles, 43*(5/6), 407–417.

Beattie, J., Spooner, F., Jordan, L., Algozzine, B., & Spooner, M. (2002). Evaluating instruction in distance learning classes. *Teacher Education and Special Education, 25*, 124–132.

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. (2012). *Student ratings of teaching: A summary of research and literature* (IDEA Paper No. 50). Manhattan: The IDEA Center.

Benton, S. L., & Pallett, W. H. (2013). Class size matters. *Inside Higher Education*. http://www.insidehighered.com/views/2013/01/29/essay-importance-class-size-higher-education. 28 Jan 2013.

Benton, S. L., Webster, R., Gross, A. B., & Pallett, W. (2010a). *IDEA technical report no. 16: An analysis of IDEA student ratings of instruction using paper versus online survey methods*. Manhattan: The IDEA Center.

Benton, S. L., Webster, R., Gross, A. B., & Pallett, W. (2010b). *IDEA technical report no. 15: An analysis of IDEA student ratings of instruction in traditional versus online courses*. Manhattan: The IDEA Center.

Benton, S. L., Duchon, D., & Pallett, W. H. (2011). Validity of self-report student ratings of instruction. *Assessment and Evaluation in Higher Education., 38*, 377–389.

Benton, S. L., Brown, R., & Li., D. (2012a). *Replication of IDEA Technical Report No. 12 tables: 2012 IDEA student ratings dataset*. Unpublished manuscript.

Benton, S. L., Gross, A., & Brown, R. (2012b, October). *Which learning outcomes and teaching methods are instructors really emphasizing in STEM courses?* Presentation at American Association of Colleges and Universities Network for Academic Renewal, Kansas City.

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.

Beran, T., Violato, C., Kline, D., & Frideres, J. (2007). What's the "use" of student ratings of instruction for administrators? *Canadian Journal of Higher Education, 37*, 27–43.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*, 48–62.

Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology, 57*, 195–203.

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.

Brener, N. D., Billy, J. O. G., & Grady, W. R. (2003). Assessment of factors affecting the validity of self-reported health-risk behavior among adolescents: Evidence from the scientific literature. *Journal of Adolescent Health, 33*, 436–457.

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

Brown, M. (2012, July). Learning analytics: Moving from concept to practice. In *Educause Learning Initiative Brief*. Retrieved from the EDUCAUSE website: http://net.educause.edu/ir/library/pdf/ELIB1203.pdf

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.

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.

Burgoon, J. K., Birk, T., & Pfau, M. (1990). Nonverbal behaviors, persuasion, and credibility. *Human Communication Research, 17*, 140–169.

Campbell, J., & Bozeman, W. (2007). The value of student ratings: Perceptions of students, teachers, and administrators. *Community College of Research and Practice, 32*(1), 13–24.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Carrell, S., & West, J. (2010). Does professor quality matter? Evidence from random assignments of students to professors. *Journal of Political Economy, 118*(3), 409–432.

Carrier, N. A., Howard, G. S., & Miller, W. G. (1974). Course evaluations: When? *Journal of Educational Psychology, 66*, 609–613.

Cashin, W. E (1989). *Defining and evaluating college teaching* (IDEA Paper No. 21). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.
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: Kansas State University, Center for Faculty Evaluation and Development.
Cashin, W. E. (2003). Evaluating college and university teaching: Reflections of a practitioner. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 531–593). Dordrecht: Kluwer Academic Publishers.
Cashin, W. E., & Downey, R. G. (1992). Using global student ratings for summative evaluation. *Journal of Educational Psychology, 84*, 563–572.
Centra, J. A. (1976). The influence of different directions on student ratings of instruction. *Journal of Educational Measurement, 13*, 277–282.
Centra, J. A. (1979). *Determining faculty effectiveness: Assessing teaching, research, and service for personnel decisions and improvement*. San Francisco: Jossey-Bass.
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: 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. (2009). Student evaluation of teaching: Are they related to what students learn? *Journal of Marketing Education, 31*, 16–30.
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.
Cooper, A. M., Sobell, M. B., Sobell, L. C., & Maisto, S. A. (1981). Validity of alcoholics' self-reports: Duration data. *International Journal of Addiction, 16*, 401–406.
Costin, F. (1968). A graduate course in the teaching of psychology: Description and evaluation. *Journal of Teacher Education, 19*, 425–432.
Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: reliability, validity, and usefulness. *Review of Educational Research, 41*, 511–535.
Creasman, P. A. (n.d.). *IDEA Paper No. 52: Considerations in online course design*. Manhattan: The IDEA Center.
Davis, B. G. (2009). *Tools for teaching* (2nd ed.). San Francisco: Jossey-Bass.
Del Boca, F. K., & Noll, J. A. (2002). Truth or consequences: The validity of self-report data in health services research on addictions. *Addiction, 95*, 347–360.
Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment and Evaluation in Higher Education, 29*, 611–623.
Dorn, D. S. (1987). The first day of class: Problems and strategies. *Teaching Sociology, 15*, 61–72.
Dyckhoff, A. L. (2011). Implications for learning analytics tools: A meta-analysis of applied research questions. *International Journal of Computer Information Systems and Industrial Management Application, 3*, 594–601.

Erdle, S., Murray, H. G., & Rushton, J. P. (1985). Personality, classroom behavior, and student ratings of college teaching effectiveness: A path analysis. *Journal of Educational Psychology, 77*, 394–407.

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

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

Feldman, K. A. (1976b). The superior college teacher from the students' view. *Research in Higher Education, 5*, 243–288.

Feldman, K. A. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education, 6*, 233–274.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education, 9*, 199–242.

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. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education, 21*, 45–116.

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. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities. *Research in Higher Education, 28*, 291–344.

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: Springer.

Forsyth, D. R. (2003). *Professor's guide to teaching: Psychological principles and practices*. Washington, DC: American Psychological Association.

Franklin, J., & Theall, M. (1992). *Disciplinary differences: Instructional goals and activities, measures of student performance, and student ratings of instruction*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Freier, M. C., Bell, R. M., & Ellickson, P. (1991). *Do teens tell the truth? The validity of self-reported tobacco use by adolescents*. Santa Monica: The Rand Corporation.

Frey, P. W. (1976). Validity of student instructional ratings as a function of their timing. *Journal of Higher Education, 47*, 327–336.

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

Grant, D. (2007). Grades as information. *Economics of Education Review, 26*, 201–214.

Greenwald, A. G., & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*, 743–751.

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.

Harrison, P. D., Douglas, D. K., & Burdsal, C. A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education, 45*, 311–323.

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

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

Hativa, N., & Raviv, A. (1996). University instructors' ratings profiles: Stability over time, and disciplinary differences. *New Directions for Teaching and Learning* No. 64. San Francisco: Jossey-Bass.

Hativa, N., Barak, R., & Simhi, E. (2001). Exemplary university teachers: Knowledge and beliefs regarding effective teaching dimensions and strategies. *Journal of Higher Education, 72*, 699–729.

Hativa, N., Many, A., & Dayagi, R. (2010). The whys and wherefores of teaching evaluation by their students [Hebrew]. *Al Hagova, 9*, 30–37.

Hobson, S. M., & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching, 49*, 26–31.

Hornbeak, J. L. (2009). *Teaching methods and course characteristics related to college students' desire to take a course.* K-State electronic theses, dissertations, and reports: 2004. http://hdl.handle.net/2097/1367

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.

Hoyt, D. P., & Cashin, W. E. (1977). *IDEA technical report no. 1: Development of the IDEA system*. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Hoyt, D. P., & Lee, E. (2002a). *Technical report no. 12: Basic data for the revised IDEA system*. Manhattan: The IDEA Center.

Hoyt, D. P., & Lee, E. J. (2002b). *Technical report #13: Disciplinary differences in student ratings*. Manhattan: Kansas State University, IDEA Center.

Hoyt, D. P., & Pallett, W. H. (n.d.). *IDEA paper no. 36, appraising teaching effectiveness: Beyond student ratings*. Manhattan: The IDEA Center.

Huston, T. (2005). *Research report: Race and gender bias in student evaluations of teaching*. Retrieved April 16, 2013, 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.

Kember, D., & Leung, D. Y. P. (2011). Disciplinary differences in student ratings of teaching quality. *Research in Higher Education, 52*, 278–299.

Kember, D., McKay, J., Sinclair, K., & Wong, F. K. Y. (2008). A four-category scheme for coding and assessing the level of reflection in written work. *Assessment and Evaluation in Higher Education, 33*(4), 363–379.

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

Krathwohl, D. R. (1998). *Methods of educational and social science research*. New York: Longman.
Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* (New directions for institutional research, no. 109, pp. 9–25). San Francisco: Jossey-Bass.
Kulik, J. A., & McKeachie, W. J. (1975). The evaluation of teachers in higher education. In F. N. Kerlinger (Ed.), *Review of research in education* (Vol. 3, pp. 210–240). Itasca: F. E. Peacock.
Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75*, 63–82.
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.
Leventhal, L., Abrami, P. C., Perry, R. P., & Breen, L. J. (1975). Section selection in multi-section courses: Implications for the validation and use of teacher rating forms. *Educational and Psychological Measurement, 35*, 885–895.
Li, Y. (1993). *A comparative study of Asian and American students' perceptions of faculty teaching effectiveness at Ohio University.* Unpublished doctoral dissertation, Ohio University, Athens.
Linse, A. R. (2012). *Faculty strategies for encouraging their students to fill out the SRTEs*. Retrieved April 16, 2013, from http://www.schreyerinstitute.psu.edu/IncreaseSRTERespRate/
Marincovich, M. (1999). Using student feedback to improve teaching. In P. Seldin & Associates (Eds.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 45–69). Bolton: Anker.
Marks, R. B. (2000). Determinants of student evaluation of global measures of instructor and course value. *Journal of Marketing Education, 22*(2), 108–119.
Marsh, H. W. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology, 74*, 264–279.1.
Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707–754.
Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253–388.
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: Springer.
Marsh, H. W., & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education, 64*, 1–17.
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. (1979). Long-term stability of students' evaluations: A note on Feldman's consistency and variability among college students in rating their teachers and courses. *Research in Higher Education, 10*, 139–147.
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.

Marsh, H. W., & Ware, J. E. (1982). Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. *Journal of Educational Psychology, 74*, 126–134.

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.

McCarthy, M. A., Niederjohn, D. M., & Bosack, T. N. (2011). Embedded assessment: A measure of student learning and teaching effectiveness. *Teaching of Psychology, 38*(2), 78–82.

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., & 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.

McPherson, M. A., & Todd Jewell, R. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly, 88*(3), 868–881.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Old Tappan: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 1995*(50), 741–749.

Midanik, L. (1988). Validity of self-report alcohol use: A literature review and assessment. *British Journal of Addictions, 83*, 1019–1030.

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

Murray, H. G. (1997). Effective teaching behaviors in the college classroom. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 171–204). New York: Agathon Press.

Murray, H. G. (2005, June). *Student evaluation of teaching: Has it made a difference?* Paper presented at the annual meeting for the Society of Teaching and Learning in Higher Education, Charlottetown, Prince Edward Island.

Murray, H. G. (2007). Low-inference teaching behaviors and college teaching effectiveness: Recent developments and controversies. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 145–200). Dordrecht: Springer.

Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology, 82*, 250–261.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education, 48*, 630–635.

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: Jossey-Bass.

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. 5, pp. 27–44). San Francisco: Jossey-Bass.

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.

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

Pallett, W. H. (2006). Uses and abuses of student ratings. In P. Seldin (Ed.), *Evaluating faculty performance* (pp. 50–65). Bolton: Anker Publishing Company, Inc.

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.

Patrick, D. L., Cheadle, A., Thompson, D. C., Diehr, P., Koepsell, T., & Kinne, S. (1994). The validity of self-reported smoking: A review and meta-analysis. *American Journal of Public Health, 84*(7), 1086–1093.

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

Perry, R. P., & Smart, J. C. (Eds.). (1997). *Effective teaching in higher education: Research and practice*. New York: Agathon Press.

Perry, R. P., & Smart, J. C. (Eds.). (2007). *The Scholarship of teaching and learning in higher education: An evidence-based perspective*. Dordrecht: Springer.

Perry, R. P., Niemi, R. R., & Jones, K. (1974). Effect of prior teaching evaluations and lecture presentation on ratings of teaching performance. *Journal of Educational Psychology, 66*, 851–856.

Perry, R. P., Abrami, P. C., & Leventhal, L. (1979a). Educational seduction: The effect of instructor expressiveness and lecture content on student ratings and achievement. *Journal of Educational Psychology, 71*, 107–116.

Perry, R. P., Abrami, P. C., Leventhal, L., & Check, J. (1979b). Instructor reputation: An expectancy relationship involving student ratings and achievement. *Journal of Educational Psychology, 71*, 776–787.

Ray, J. J. (1987). The validity of self-reports. *Personality Study and Group Behaviour, 1*, 68–70.

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

Renaud, R. D., & Murray, H. G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education, 46*, 929–953.

Schmelkin, L. P., Spencer, K. J., & Gellman, E. S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education, 38*, 575–592.

Schulze, E., & Tomal, A. (2006). The chilly classroom: Beyond gender. *College Teaching, 54*(3), 263–270.

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.

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.

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.

Sullivan, A. M., & Skanes, G. R. (1974). Validity of student evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology, 66*(4), 584–590.

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

The IDEA Center. (2008). Best practices for online response rates. Retrieved April 16, 2013, from http://www.theideacenter.org/OnlineResponseRates

Theall, M., & Feldman, K. A. (2007). Commentary and update on Feldman's (1997) "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. 130–143). Dordrecht: Springer.

Venette, S., Sellnow, D., & McIntire, K. (2010). Charting new territory: Assessing the online frontier of student ratings of instruction. *Assessment and Evaluation in Higher Education, 35*, 101–115.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education, 23*, 191–211.

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, J. E., & Williams, R. G. (1975). The Dr. Fox effect: A study of lecture effectiveness and ratings of instruction. *Journal of Medical Education, 50*, 149–156.

Weimer, M. (2009). Teachers who improved. *The Teaching Professor, 23*, 2.

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *Journal of Economic Education, 40*(3), 227–261.

Williams, R. G., & Ware, J. E. (1976). Validity of student ratings of instruction under different incentive conditions: A further study of the Dr. Fox effect. *Journal of Educational Psychology, 68*, 48–56.

Williams, R. G., & Ware, J. E. (1977). An extended visit with Dr. Fox: Validity of student ratings of instruction after repeated exposure to a lecturer. *American Educational Research Journal, 14*, 449–457.

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

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.