1

This chapter reviews the conclusions on which most experts agree, cites some of the main sources of support for these conclusions, and discusses some dissenting opinions and the research support for those opinions.

Student Ratings: Validity, Utility, and Controversy

James A. Kulik

Student ratings are an old topic in higher education. Seventy-five years have passed since students at the University of Washington filled out what were arguably the first student rating forms (Guthrie, 1954). Almost as long a time has passed since researchers at Purdue University published the first research studies on student ratings (Remmers and Brandenburg, 1927). But student ratings are not yet a stale topic. Teachers still talk about them, researchers still study them, and most important, students still fill out the forms—millions of them every year—in college classes throughout the country.

Seldin's surveys on teaching evaluation (1993a) show just how widespread rating systems have become. About 29 percent of American colleges reported using student ratings to evaluate teaching in Seldin's 1973 survey, 68 percent of colleges reported using them in his 1983 survey, and 86 percent reported using them in his 1993 survey. Seldin reported that no other data source gets more attention in the evaluation of teaching—not classroom visits, not examination scores, and not self-reports.

Rating results are also being used today in more ways than ever before. Colleges originally set up rating systems to serve two purposes: to help administrators monitor teaching quality and to help teachers improve their teaching (Guthrie, 1954). Today, ratings serve many purposes. At my own institution, administrators and administrative committees use ratings in hiring new faculty, in annual reviews of current faculty, in promotion and tenure decisions, in school accreditation reviews, in selecting faculty and graduate students for teaching awards and honors, and in assigning teachers to courses. Faculty members use ratings when trying to improve their teaching effectiveness, in documenting their effectiveness internally and externally,

and in monitoring the performance of their graduate student assistants. Graduate student instructors use ratings in developing their teaching skills and in documenting these skills in job applications. Student groups use the ratings in selecting courses and in selecting teachers for awards and honors.

Many teachers applaud the increased use of ratings on college campuses. They view ratings as reliable and valid measures that bring scientific accuracy to the evaluation of teaching, and they also argue that ratings give students more of a voice in their education. But not everyone is so enthusiastic. Some teachers view ratings as meaningless quantification. They fear that students too often use the power of their pencils to get even with professors and warn that rating systems may turn the evaluation of effective teaching into a personality contest.

Researchers have collected a wealth of data on student ratings over the years. One might suppose that the research studies on ratings are similar to many other studies in education: conflicting, confusing, and inconclusive. And some of the studies of ratings are. It is a mistake, however, to ignore this research literature.

In this chapter, I review the conclusions on which most experts agree. I cite some of the main sources of support for these conclusions, and I discuss some dissenting opinions and the research support for those opinions.

Validity of Student Ratings

To say that student ratings are valid is to say that they reflect teaching effectiveness. It would therefore seem to be a straightforward matter to assess the validity of student ratings. All we have to do is to correlate student ratings with teaching effectiveness scores. If ratings are valid, students will give good ratings to effective teachers and poor ratings to ineffective ones. The size of the correlation between ratings and effectiveness will provide a precise index of the validity of student ratings.

The catch is that no one knows what measure to use as *the* criterion of teaching effectiveness. Researchers have long searched for the perfect criterion. Among the criteria that they have examined are measures of student learning, alumni judgments of teachers, and classroom observations by experts. But the search has proved futile because each of these criteria is far from perfect.

Scriven (1983) is especially clear about the shortcomings of these measures. About learning measures, he has written, "The best teaching is not that which produces the most learning" (p. 248). According to Scriven, good examination performance may result from a number of factors besides good teaching. For example, a teacher may put so much pressure on students that they abandon their work in other classes. Such pressure tactics may produce good examination scores in the teacher's course, but the students pay too high a price for their accomplishments. The occasional use of such tactics

by college teachers illustrates the general point that good examination performance can result from unethical or bad teaching. On a more practical note, the tests that teachers administer are often far from perfect. They are sometimes neither reliable nor valid measures of what is learned in class. They are usually an inadequate indicator of the influence that great teachers have on students' lives.

Scriven (1983) is equally clear about the weaknesses in expert visits to a classroom. Using such visits to evaluate teaching, he says, "is not just incorrect, it is a disgrace" (p. 251). The visits themselves alter teaching, Scriven points out, and the number of experts and visits is usually too small to yield a reliable measure of teaching effects. Furthermore, the experts who provide the ratings usually have biases that can skew their observations. Finally, classroom talk, which is the thing that the experts observe, is only a small part of what constitutes college teaching. Many things not observable in classroom discourse are necessary for good teaching, including fair grades and valid tests.

Scriven (1983) warns that alumni surveys are "essentially useless for evaluation of teachers" (p. 254). They usually have extremely low response rates and relate to "ancestors" of current performance. Alumni perspectives are sometimes dated—teachers change and times change—and alumni views about what will be valuable for a new generation of graduates may be wrong. Scriven concedes, however, "These reasons do not exclude some use of alumni surveys in selecting Distinguished (Elderly) Teacher Awards" (p. 254).

Not all experts on ratings are as passionate about the shortcomings of these measures as Scriven is. But all experts agree with him on the practical impossibility of finding a single perfect criterion of teaching effectiveness. With such a measure, we could calculate a predictive validity coefficient for student ratings. The correlation coefficient between ratings and effectiveness would give the degree to which we could predict effectiveness from ratings. Without a perfect criterion, it is impossible to reduce the validity of student ratings to a single number.

Given this difficulty, most researchers on ratings have adopted what is sometimes called a "construct validation approach" to student ratings. This approach requires researchers to show that ratings correlate to a satisfactory degree with other admittedly partial and imperfect measures of effectiveness. Experts do not expect perfect agreement between ratings and such imperfect measures, but they do expect student ratings to correlate at least moderately with these other measures.

In this section, I focus on the agreement between ratings and four of the most credible of the indicators of effectiveness: student learning, student comments, alumni ratings, and ratings of teaching by outside observers. I conclude that rating results agree adequately, but not perfectly, with results from each of these indicators. Teachers who come out high on one measure usually come out high on other measures, too.

Students Learn More from Highly Rated Teachers. The best data on the correlation between ratings and learning come from dozens of studies of student ratings in multisection college courses. In these studies, instructors teach a section of a course in their own way, but all instructors cover the same content and administer the same common final examination. To determine whether superior ratings go with better or poorer exam performance, researchers correlate section averages on the examination with section averages on the rating scales. Researchers conducting such studies have usually found that examination and rating averages correlate positively. They have concluded therefore that students generally give high ratings to teachers from whom they learn most, and they generally give low ratings to teachers from whom they learn least.

Some of the many reviews are narrative in form (Costin, Greenough, and Menges, 1971; Kulik and McKeachie, 1975). Others are meta-analytic reports (Cohen, 1981, 1982; Feldman, 1989c; McCallum, 1984). The meta-analytic reviews are clearer than the narrative ones, and the clearest of the meta-analyses, Peter Cohen's classic report (1981), which was the first one published.

Cohen's meta-analysis covered data from forty-one studies that reported on sixty-eight separate multisection courses. Like other meta-analysts, Cohen located his studies in objective computer searches of library databases. He then expressed the outcomes of all studies in terms of product-moment correlation coefficients, and he also coded the features of the studies in quantitative or quasi-quantitative terms. Finally, Cohen calculated the average result in all studies and in various subgroups of the studies.

Cohen found a strong relationship between student ratings and student learning in the average study. The average correlation of examination score with overall rating of the teacher was .43. The average correlation of examination score with an overall rating of the course was .47. Although there is no definite standard for interpreting size of correlation, Jacob Cohen (1977) has provided some rough guidelines stating that a coefficient of about .50 is large, a coefficient of about .30 is moderate, and a correlation of about .10 is small. According to these standards, the correlation between learning and an overall rating of the teacher or the course is moderate to high.

Although the average correlation between ratings and achievement measures was moderate to high in Cohen's analysis, he observed that not all studies produced the same results. Indeed, study results varied a great deal. Some studies reported a high positive correlation between ratings and achievement; other studies reported a negative correlation. Cohen was interested in finding some factor that might explain the variation in study findings. He examined twenty study features in his attempt to explain the variation.

He found, first of all, that the items included on a rating scale could influence study findings. The correlation between ratings and achievement was high for items involving instructor skill and for those measuring teacher

and course organization. Correlation coefficients were moderate for items on teacher rapport and feedback near zero for items dealing with course difficulty.

A few other study features seemed to influence the findings. Specifically, correlation coefficients were higher in studies in which the instructors were full-time teachers, in studies in which students knew their final grade when they rated the instructor, and in studies where achievement tests were evaluated by an external evaluator. Cohen also reported that many other study characteristics (such as random assignment, course content, and availability of pretest data) were not significantly related to study findings.

Student Ratings Agree with Student Comments. Researchers have carried out only a few studies on the agreement between student ratings and the comments that students freely make about their teachers. The findings of the available studies are so clear, however, that they are worth noting (Braskamp, Ory, and Pieper, 1981; Ory, Braskamp, and Pieper, 1980). The evidence shows that ratings correlate strongly with comments that students make about their teachers both on questionnaires and in special interviews.

The most direct evidence for this point comes from Ory, Braskamp, and Pieper's study of comments and ratings (1980). These authors focused on classes in which students filled out rating forms, wrote answers to open-ended questions about the course and teacher, and spoke to consultants about the course in group interviews. The students made their ratings on 6-point scales, and the researchers coded the students' written and interview comments on the same scales. Finally, the researchers correlated the data from the three sources.

Ory and his colleagues found a remarkable degree of consistency between student ratings and the ratings of a class derived from written and interview comments. Student ratings of the course and instructor correlated .94 and .93 with ratings derived from written comments; student ratings of the course and instructor correlated .81 and .84 with ratings derived from student comments in interviews. Along with Ory and his colleagues, I conclude from this study that the extraordinarily high correlation between comments and ratings suggests that these data sources give nearly identical pictures of teaching effectiveness.

Student Ratings Agree with Observer Ratings. Murray (1983) carried out an especially careful study of the relationship between student ratings and ratings of teaching behaviors made by trained observers. He arranged for forty-nine students in an educational psychology course to report on the teaching behaviors of fifty-four college teachers. From student ratings made in an earlier semester, these teachers could be classified as high, medium, or low in effectiveness. Six to eight of the observers rated each of the fifty-four teachers in three separate one-hour class periods; the observers thus spent a total of eighteen to twenty-four hours with each teacher. During the three-month observation period, the observers saw clear differences

in the teaching behavior of the three groups. In all, the three groups differed on twenty-six individual behaviors. The sharpest differences were in behaviors indicating teacher clarity, enthusiasm, and rapport. Highly rated teachers were high and low-rated teachers were low in these three qualities.

Feldman (1989c) reviewed findings from Murray's study and four other studies that correlated student ratings with ratings made by outside observers. The average correlation coefficient between student ratings and observer ratings in these studies was .50. By conventional standards, this is a high correlation. It is especially impressive when we consider that outside observers and students do not have access to the same data. For example, observers are usually not aware of teacher behavior outside the classroom. They usually know little or nothing about the quality of teacher comments to students on their written work, the teacher's fairness in grading students, or the teacher's availability to students outside of class.

The essential point is that students give favorable ratings to teachers who get good marks from outside observers, and they give unfavorable ratings to teachers who get poor marks. Thus student ratings correlate highly with ratings by outside observers.

Student Ratings Agree with Alumni Ratings. The best evidence of agreement between student and alumni ratings of teachers comes from a longitudinal study by Overall and Marsh (1980). The fourteen hundred students in this study filled out end-of-term evaluation forms in all the courses they took during a three-year period. One year after the students graduated and one to four years after the students completed these courses, the students again filled out evaluation forms on their courses. The end-of-term ratings in one hundred courses correlated .83 with the follow-up ratings, and the median rating at the two times was nearly identical.

Additional support for the stability of ratings comes from cross-sectional studies. In these studies, different cohorts of students provide the current-student and alumni ratings. The cross-sectional design is weaker than a longitudinal design because the different cohorts of students base their ratings on different experiences with a teacher. Feldman (1989c) reviewed results from six cross-sectional studies. He found an average correlation coefficient of .69 between current-student and alumni ratings. By Jacob Cohen's standards (1977), this is a remarkably high correlation.

Thus current students and alumni give similar ratings to teachers. The findings do not support the argument that students can evaluate their courses only after they have been asked to apply course material in further courses or in their postgraduation pursuits. Instead, current students give favorable ratings to teachers whom alumni remember fondly and poor marks to teachers whom alumni remember unfavorably.

The central point that emerges from research studies and reviews on validity of ratings is that teachers who receive high ratings from their students receive high marks on other credible criteria of teaching effectiveness. Students give high ratings to the teachers from whom they learn most. They

also comment favorably about these teachers in writing and in interviews. In addition, outside observers give highly rated teachers excellent ratings, and alumni ratings of the teachers are excellent. Researchers have studied the agreement of student ratings and several other possible measures of teaching effectiveness, including self-ratings and ratings made by departmental colleagues who have not visited the teacher's classroom (Feldman, 1989c). I consider these to be less satisfactory measures of teaching effectiveness, and so I have not reviewed findings on such measures here. It is worth noting, however, that student ratings agree well with these measures too.

Utility of Student Ratings

Student rating programs are meant to improve college teaching in at least two ways. First, rating programs are meant to have effects at the institutional level. They may influence an institution's hiring decisions, merit increases, promotion and tenure decisions, and course assignments. Ratings may thus influence who teaches at a college, what courses they teach, and how much attention faculty members give to teaching. In addition, ratings are meant to have effects on individual teachers. Rating results give teachers information that they may use when trying to improve their own teaching.

Researchers have not yet developed a way of studying institutional effects of rating systems. They can point to the ubiquity of rating programs on college campuses or the longevity of many programs as presumptive evidence for the salutary effects of these programs, but hard data on institutional effects are scarce or nonexistent. Fortunately, researchers have paid far more attention to effects of student ratings on individual teachers. Researchers have carried out numerous studies on this topic, and reviewers agree on the main conclusions that can be drawn from the studies.

The basic design that researchers usually use to study rating effects on individual teachers is a two-group design. One group of teachers receives rating feedback in the middle of a course, and another group of teachers does not receive such feedback. At the end of the course, students again rate the teachers. To determine whether the midterm feedback from students is effective, the researcher compares the end-of-term ratings for the two groups.

Two studies carried out by Marsh and his colleagues (Marsh, Fleiner, and Thomas, 1975; Overall and Marsh, 1979) provide a good introduction to results in this area. In the first study, Marsh and his colleagues returned midterm student rating results to faculty and found that midterm feedback has a positive but modest effect. In the second study, the researchers met with instructors in the feedback group to discuss the evaluations and possible strategies for improvement. Not only did teachers in the feedback-plus-consultation group receive better end-of-term ratings from their students, but their students also performed better on the final examination.

Peter Cohen (1980) carried out a meta-analysis of findings in twenty-two studies of feedback effectiveness. His results parallel the results of the

two studies by Marsh and colleagues. Cohen found that midterm feedback alone has a modest effect on end-of-term ratings. Such feedback raised end-of-term ratings by an average of about 0.1 rating point. Cohen also found that effects of midterm feedback are greater when instructors receive some consulting help along with the midterm ratings. End-of-term ratings went up by about 0.3 rating point in these circumstances.

The picture that emerges from the literature on utility of ratings is a hopeful one, but it has many blank areas. We know that rating programs have a long history in American higher education and that rating programs are ubiquitous on college campuses today. Rating programs thus seem to serve some useful purpose, but research on the effects of rating programs on colleges and universities is almost nonexistent. Much more research is available on the effect that ratings have on individual teachers. Research studies indicate that rating feedback helps teachers improve their teaching performance. The studies also suggest that student feedback is especially useful when rating results are coupled with consultation on improvement strategies.

Another View of Ratings

Analysts who question the validity and utility of student ratings seldom cite the evidence that I have reviewed. They are more likely to cite findings of a handful of well-known studies that are critical of ratings. These studies include Rodin and Rodin's study of student ratings and learning (1972), Naftulin, Ware, and Donnelly's study of "educational seduction" and ratings (1973), Ambady and Rosenthal's study of thin slices of expressive behavior and ratings (1992), Greenwald and Gillmore's study of grading, student work, and ratings (1997), and Williams and Ceci's study of expressiveness and ratings (1997). The studies suggest that instead of measuring teaching effectiveness, ratings reflect peripheral factors, such as teacher personality or grading standards.

I shall now briefly describe and comment on the five studies and their findings.

Do High Ratings Imply Low Learning? Rodin and Rodin (1972) reported a negative correlation of –.75 between student rating and student learning measures in an undergraduate calculus course. They concluded from this study that students rate most highly instructors from whom they learn least, and they rate least favorably instructors from whom they learn most. The Rodin and Rodin report on this research appeared in the prestigious and widely read journal *Science*. There it probably attracted more attention than any study of ratings ever had before. Critics of ratings still sometimes cite the Rodin and Rodin finding as evidence that student ratings lack validity.

Experts on ratings, however, have roundly criticized the study (for example, Doyle, 1975; Marsh, 1984). Critics point out that the ratings collected in the study were not of the course instructor but rather of the

instructor's eleven teaching assistants. These teaching assistants actually played a minor role in course instruction. In addition, the learning measure was not a test given under standard conditions at the end of the course. Instead, Rodin and Rodin measured student learning by counting the number of examination problems that a student was able to solve during the term. The researchers gave students a total of forty examination problems, one after each course unit, and they allowed students who did not solve a problem on the first attempt to try again as many as six times without penalty. Furthermore, Rodin and Rodin had each teaching assistant score the problems for his or her own students. Differences among teaching assistants in grading standards were thus confounded with differences among them in teaching performance. Marsh and Doyle have both speculated about how these unique features of the Rodin and Rodin study could produce a spurious negative correlation between ratings and learning, and they have concluded that the methodological flaws of the study make it a poor basis for drawing conclusions about ratings and learning.

There is a more important reason for questioning the Rodin and Rodin conclusions. Their findings are an anomaly; their results are an outlier in the literature on ratings and learning. As I have already pointed out, Peter Cohen (1981) has written an authoritative and comprehensive review of the literature on ratings and learning. He found forty-one studies (including the Rodin and Rodin study) that reported on the correlation between ratings and achievement in a total of sixty-eight courses. The correlation coefficient in the Rodin and Rodin course was −.75. The average correlation coefficient between instructor rating and learning in all the courses was .43. No other study reported a correlation coefficient as low as the one found by Rodin and Rodin.

For decades, experts on ratings have been writing epitaphs for the Rodin and Rodin study, but it has refused to go away. In 1975, Doyle wrote, "To put the matter bluntly, the attention received by the Rodin and Rodin study seems disproportionate to its rigor, and their data provide little if any guidance in the validation of student ratings" (p. 59). In 1984, Marsh wrote, "In retrospect, the most interesting aspect of this study was that such a methodologically flawed study received so much attention" (p. 720). Today, twenty-five years after Rodin and Rodin published their article, we can do nothing better than to look at their findings in context. The rule is that students rate most highly teachers from whom they learn most. The Rodin and Rodin study may be the exception that proves the rule.

Do Ratings Measure Showmanship? In what has come to be known as the "Dr. Fox study," a trained actor, introduced as Dr. Fox, delivered a lecture on mathematical game theory to a group of medical educators (Naftulin, Ware, and Donnelly, 1973). Dr. Fox presented incorrect information, cited nonexistent references, and used neologisms as basic terms. Nonetheless, the great majority of Dr. Fox's audience rated his lecture favorably. The study produced a term that is still heard in discussions of ratings: "the Dr. Fox effect."

The term refers to the use of an entertaining style to "seduce" students into giving favorable evaluations to a teacher who is weak on content. The term suggests that student ratings reflect style rather than substance.

Critics of ratings have seized on this study as strong evidence for the invalidity of student ratings, but rating experts are quick to point out that the study has many methodological flaws (Abrami, Leventhal, and Perry, 1982; Frey, 1979; Marsh and Ware, 1982). Frey, for example, writes that "this study represents the kind of research that teachers make fun of during the first week of an introductory course in behavioral research methods. Almost every feature of the study is problematic" (p. 1).

The most serious charge leveled against the Dr. Fox study is irrelevance. Dr. Fox's lecture and his audience's reaction to it are a far cry from college teaching and student ratings. For example, Dr. Fox gave only one lecture before being rated. In college courses, students base their ratings on numerous lectures, the course outline, the reading material, testing, and grading. Dr. Fox might have bamboozled his audience during a single lecture, but surely everyone would have caught on to the fraud if Dr. Fox were the lecturer in a semester-long college course. In addition, Dr. Fox lectured on a topic that was completely unknown to his audience. Students in most college courses are not completely ignorant of the subject matter of the class. We can be sure that Dr. Fox would have received quite different ratings had he delivered his lecture to upper-division undergraduate or graduate students in mathematics.

My essential point is that the Dr. Fox paradigm does not apply to student ratings of college teaching. We may be able to draw conclusions about the gullibility of medical educators from the study, but surely we should not let Dr. Fox (or his creators) seduce us into drawing conclusions about student ratings. The Dr. Fox experiment is fundamentally irrelevant to student ratings of college teaching.

Do Ratings Measure Body Language? Critics of ratings sometimes cite Ambady and Rosenthal's findings (1992) as proof that student ratings are superficial. These researchers investigated what they call "thin slices of expressive behavior." These are very brief observations from which observers form impressions of others. In Ambady and Rosenthal's study, observers who saw only thirty-second silent video clips of teachers could predict the end-of-course ratings of the teachers quite accurately. The correlation between the observer and student ratings was .76. It is worth noting that Ambady and Rosenthal considered the end-of-course ratings to be a sound criterion of teaching quality, and they therefore concluded that observers can form surprisingly accurate impressions of others based on the briefest of observations.

Critics of ratings have drawn a different conclusion. If a complete stranger can guess a teacher's end-of-course ratings after viewing only a soundless thirty-second video clip of the teacher, they ask, what do end-of-course ratings actually measure? Can ratings possibly be measuring any-

thing important? Is it not more likely that end-of-course ratings reflect only superficial expressive behavior?

It is important to note that Ambady and Rosenthal's study was a very small study that involved only thirteen teachers. The correlation of .76 between observer and student ratings must therefore have a large standard error. The true correlation between the two variables could thus fall anywhere within a range almost one-half-point wide. My own guess is that the true correlation is near the lower of these values. I base this guess on Feldman's review of the literature on agreement between student end-of-course ratings and ratings made by expert observers (1989c). The studies that Feldman reviewed involved longer observation periods and the observers not only saw what teachers were doing but also heard what teachers were saying. In Ambady and Rosenthal's words, these experts observed "thick slices of behavior." The average correlation between these thick slices of behavior and student ratings was .50. Common sense suggests that thick slices of teaching behavior will predict end-of-course ratings much better than thin slices do. I would therefore expect most researchers to find correlation coefficients between ratings and thin slices of behavior to be considerably below .50.

Do High Ratings Reflect Lenient Grading? Greenwald and Gillmore (1997) analyzed the agreement between measures of student effort, the grades that student expect in their classes, and their ratings of these classes. They concluded from their analyses that grading leniency exerts an important influence on both student ratings and student effort. They also concluded that student rating results should always be statistically adjusted to remove the unwanted influence of grading leniency.

Greenwald and Gillmore's data came from two hundred undergraduate courses at the University of Washington. The researchers found a positive correlation between student ratings of teachers and the grades given out by the teachers. Specifically, they found that teachers who get high ratings from students tend to give out higher grades, whereas teachers who get low ratings tend to give out lower grades. Greenwald and Gillmore also found a negative relationship between the grades given out in a course and the amount of work students do for the course. Specifically, they found that students reported working harder in classes where professors generally gave low grades and slacked off in classes in which professors gave high grades.

Other researchers have studied these same variables and have made several points about Greenwald and Gillmore's results. First, correlation coefficients between these variables tend to be small. Researchers typically find a correlation of about .2 between grades and ratings. Researchers also find a small correlation between student effort and ratings, but the correlation seems to be a function of the way student effort is measured. The correlation between effort and grading leniency is small and positive with some measures of student effort; it is small and negative with other measures. Second, it is difficult to interpret the correlation coefficients. For example,

researchers have proposed several explanations for the correlation between grades and ratings:

- The ratings that a teacher receives might influence the teacher to be either stingy or generous with grades.
- The grades that students receive might influence them to give high or low ratings to a teacher.
- A third factor, such as good teaching, might stimulate students to perform well in a course (and thus receive high grades) and might also lead students to give the course high ratings.

Greenwald and Gillmore (1997) found that their correlation coefficients fit a model that makes grading leniency the prime influence on both ratings and student effort. The model specifies that a strict grading policy leads students to put more effort into a course, but it also leads to low ratings for the course. A generous grading policy has the opposite effects. It encourages students to slack off, but it also leads to high ratings. Greenwald and Gillmore fear that instructors, sensing the relationship between grades and ratings, may be tempted to grade higher to get higher ratings from students. One result of such lenient grading might be a decline in the amount of effort that students put into their courses. The ultimate consequence could be a "dumbing down" of college education. To prevent such a thing from happening, Greenwald and Gillmore suggest the use of a statistical correction to ratings that would remove the undesirable influence of grading leniency.

Ratings experts have questioned Greenwald and Gillmore's conclusion and their proposed statistical correction of student ratings (d'Apollonia and Abrami, 1997; Marsh and Roche, 1997; McKeachie, 1997). Among their concerns are the correlation coefficients that Greenwald and Gillmore use in their models. The correlation coefficients show the influence of the range of courses that Greenwald and Gillmore included in their analyses. It is no secret that average grades, work requirements, and ratings vary by subject in most colleges, and average grades, work requirements, and ratings also vary by course level. These factors affect the size of correlation coefficients between student efforts, grades, and ratings, and Greenwald and Gillmore should have removed their influence from their correlation coefficients. A further concern of the experts who reviewed Greenwald and Gillmore's study were the path models that they tested. The experts pointed out that Greenwald and Gillmore did not test a sufficient number of alternative models and that they gave too little attention to teaching effectiveness in their models.

Experts who have written about Greenwald and Gillmore's work find it stimulating but remain unconvinced by the conclusions they reached. The data on which Greenwald and Gillmore built their model seem weak, the model itself seems arbitrary, and the conclusions seem questionable. Nonetheless,

their study will have a positive influence if it inspires researchers to explore in depth the tangled web of relationships that produce significant correlation coefficients between grades, ratings, and course workload.

Do Ratings Measure Vocal Expressiveness? Williams and Ceci (1997) studied the effects that stylistic changes can have on a teacher's effectiveness. They found that changes in vocal expressiveness produced large, across-the-board increases in one teacher's student ratings, but the changes had no effect on examination scores. The researchers argued from these findings that student ratings must therefore be invalid as measures of teaching effectiveness.

Williams and Ceci's study involved a single course and a single teacher. The course was Developmental Psychology at Cornell University, and the teacher of the course was Ceci himself. In the fall term, Ceci gave the course in its usual way, the same way he had been teaching the course for twenty years. Ceci's students gave low ratings to the course on most rating scales, and they rated the course especially low on instructor enthusiasm. Ceci's rating on enthusiasm was around 2 on a 5-point scale. The university then invited Ceci to attend a workshop on teaching skills. In the workshop, Ceci was encouraged to be more expressive when lecturing.

When Ceci presented the same course content and material the next semester, he varied his vocal pitch and used more gestures in order to be more expressive. Enthusiasm appears to be exactly what Ceci's lectures needed. After incorporating the workshop suggestions into his lectures, Ceci saw his rating on enthusiasm zoom up, and his other ratings tagged along. Examination scores, however, did not go up at all. Many people might take Ceci's testimonial to be a great success story, but Williams and Ceci present it as a cautionary tale:

> Our point is not especially that content-free stylistic changes can cause students to like a course more or less; nor is it that students' general affect toward a course influences their ratings of multiple aspects of the course and its instructor (halo effects). What is most meaningful about our results is the magnitude of the changes in students' evaluations due to a content-free stylistic change by the instructor, and the challenge this poses to widespread assumptions about the validity of student ratings (p. 22).

To Williams and Ceci, three findings point to the invalidity of ratings. First, they think that Ceci's ratings changed too dramatically in response to the small changes that he made in voice and gesture. Ceci's score on enthusiasm, for example, went up more than 2 points on a 5-point scale. Second, the rating changes were across-the-board. In addition to enthusiasm, ratings went up on scales measuring amount learned, fairness of grading, and quality of the textbook. Third, rating changes were not accompanied by changes in exam scores, which seem to the authors to be the real measure of good teaching. These are important points, and each is worthy of comment.

First, changes of the magnitude that Ceci observed in his rating are exceptional. Most teachers cannot expect to profit as much as Ceci did from instructional diagnosis and consultation. As I have already pointed out, Cohen (1980) reports that the typical teacher gains only about 0.3 point from feedback and consultation. In contrast, Ceci's rating on enthusiasm went up 2 points. Ceci's lectures apparently suffered from a very definite problem, and he was fortunate that he received exactly the consultation and training he needed to overcome the communication problem.

Second, ratings went up not only for enthusiasm but on other rating scales as well. The mean ratings for the instructor and the course went up about 1 point on a 5-point scale, and even the rating of the textbook went up by 1 point (although Ceci did not change the course text). Williams and Ceci apparently expect student ratings to be more analytical and focused. If only one factor in a course changes, only one rating scale should change. Perhaps, but rating scales reflect the way people feel as well as the way they think, and feelings are often diffuse and unanalytical. For this reason, evaluation experts usually advise teachers with low ratings to concentrate on their greatest relative weakness. Fix it, the experts advise, and the whole profile of ratings may go up. Most teachers should not expect to experience as dramatic a change in rating profile as Ceci experienced, of course, but changes in profile elevation are commonplace with highly intercorrelated rating scales.

Third, ratings but not examination scores rose in Ceci's class. This presents a problem if one assumes that ratings are valuable primarily as a surrogate for examination performance. Scriven (1983) argues that this assumption is unjustified. Learning and ratings are connected, he warns, but not in a simple way. As I have already pointed out, there is ample evidence that students generally learn more from teachers who get high ratings, but the relationship between examination performance and student ratings is not perfect. Other factors than teaching effectiveness affect student performance on examinations. Bad teachers sometimes put unreasonable pressures on students, and that unethical behavior may produce maximum exam scores. Great teachers sometimes influence students in ways that examinations can never measure.

My own impression is that Williams and Ceci have not given ratings their due. I think that ratings brought important benefits to Ceci, his students, and his university. Although Ceci's current students are not doing better on his tests than his past students did, his current students have positive attitudes toward their teacher and his course. The attitudes of his past students were negative and critical. To me, this change in student attitudes does not seem trivial.

The studies that I have reviewed here challenge the expert consensus on rating validity and utility. They suggest that student ratings do not reflect teaching effectiveness. Instead, ratings seem to reflect factors that are irrelevant or antithetical to good teaching. In the case of Rodin and Rodin's

study (1972), ratings seem to indicate low teacher standards. For the authors of the Dr. Fox studies (Naftulin, Ware, and Donnelly, 1973), ratings measure showmanship. Ambady and Rosenthal's findings (1992) suggest that ratings measure little more than body language. For Greenwald and Gillmore (1997), grading leniency leads to good ratings. And for Williams and Ceci (1997), variation in vocal pitch and gestures make all the difference between good and poor ratings.

There are flaws in each of these five studies. The flaws are clearest in the studies by Rodin and Rodin and by Naftulin and his colleagues. In fact, the flaws in these two studies are so deep that most experts dismiss the findings of the studies as largely irrelevant. It is too soon to know whether the studies by Ambady and Rosenthal, Greenwald and Gillmore, and Williams and Ceci will suffer the same fate. It is true that the findings of these studies are anomalous and that experts have challenged the study findings on methodological grounds. Nonetheless, the final word has not been written on these studies. We need follow-up work on the issues they raise so that we can judge how dependable their findings are.

Conclusion

The vast majority of the colleges in this country now use student ratings to evaluate teaching, and at some colleges, rating systems have been in use for decades. It seems unlikely, therefore, that student ratings are going to disappear from college campuses anytime soon. If anything, the trend seems to be toward an increasing use of student ratings in higher education.

Given the ubiquity and longevity of rating systems in colleges, we should be grateful that a research base exists from which we can draw conclusions about the validity and utility of ratings. Guthrie and Remmers initiated the research tradition in the 1920s, and it is still alive today. Researchers continue to carry out original studies of ratings, and reviewers continue to write reviews that interpret the findings.

What do the research studies show? First, the studies show that student ratings agree well with other measures of teaching effectiveness: learning measures, student comments, expert observations, and alumni ratings. The correlation between student ratings and examination scores and between ratings and classroom observations is high. Second, research studies also show how useful ratings can be to teachers. The studies show that teachers profit from the information that ratings provide. They profit from ratings alone, and they profit even more from rating results accompanied by instructional consultation. Ratings alone raise teaching effectiveness scores a little. Ratings plus consultation raise effectiveness more.

In addition to yea-sayers, student ratings research has its nay-sayers. These are the researchers who are critical of student ratings and student ratings research. The nay-sayers have actually contributed a good deal of vitality to ratings research. In the 1970s, for example, Rodin and Rodin (1972)

shook up the experts with their study on ratings and learning, and Naftulin and his colleagues (1973) further stirred up things with their Dr. Fox study. The unexpected results that emerged in the Rodin and Rodin study stimulated researchers to write authoritative reviews on the topic of ratings and learning, and the Dr. Fox study stimulated researchers to carry out a series of studies on educational seduction. More recently, researchers have presented challenging findings on the influence on ratings of body language, grading leniency, and variety in vocal pitch. I hope that researchers will respond to the challenge of these recent studies by attempting to replicate and build on their findings.

References

Abrami, P. C., Leventhal, L., and Perry, R. P. "Educational Seduction." *Review of Educational Research,* 1982, *52,* 446–464.

Ambady, N., and Rosenthal, R. "Half a Minute: Predicting Teacher Evaluations from Thin Slices of Nonverbal Behavior and Physical Attractiveness." *Journal of Personality and Social Psychology,* 1992, *64,* 431–441.

Braskamp, L. A., Ory, J. C., and Pieper, D. M. "Student Written Comments: Dimensions of Instructional Quality." *Journal of Educational Psychology,* 1981, *73,* 65–70.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* (rev. ed.) Orlando, Fla.: Academic Press, 1977.

Cohen, P. A. "Effectiveness of Student-Rating Feedback for Improving College Instruction: A Meta-Analysis." *Research in Higher Education,* 1980, *13,* 321–341.

Cohen, P. A. "Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies." *Review of Educational Research,* 1981, *51,* 281–309.

Cohen, P. A. "Validity of Student Ratings in Psychology Courses: A Research Synthesis." *Teaching of Psychology,* 1982, *9,* 78–82.

Costin, F., Greenough, W. T., and Menges, R. J. "Student Ratings of College Teaching: Reliability, Validity, and Usefulness." *Review of Educational Research,* 1971, *41,* 511–536.

d'Apollonia, S., and Abrami, P. C. "Navigating Student Ratings of Instruction." *American Psychologist*, 1997, *52*(11), 1198–1208.

Doyle, K. O. *Student Evaluation of Instruction.* Lexington, Mass.: Heath, 1975.

Feldman, K. A. "An Afterword for '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,* 1989a, *31,* 315–318.

Feldman, K. A. "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,* 1989b, *30,* 583–645.

Feldman, K. A. "Instructional Effectiveness of College Teachers as Judged by Teachers Themselves, Current and Former Students, Colleagues, Administrators and External (Neutral) Observers." *Research in Higher Education,* 1989c, *30,* 137–194.

Frey, P. W. "The Dr. Fox Effect and Its Implications." *Instructional Evaluation,* 1979, *3,* 1–5.

Greenwald, A. G., and Gillmore, G. M. "Grading Leniency Is a Removable Contaminant of Student Ratings." *American Psychologist,* 1997, *52,* 1209–1217.

Guthrie, E. R. *The Evaluation of Teaching: A Progress Report.* Seattle: University of Washington, 1954.

Kulik, J. A., and McKeachie, W. J. "The Evaluation of Teachers in Higher Education." In F. N. Kerlinger (ed.), *Review of Research in Education.* Vol. 3. Itasca, Ill.: Peacock, 1975.

Marsh, H. W. "Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility." *Journal of Educational Psychology,* 1984, *76,* 707–754.

Marsh, H. W., Fleiner, H., and Thomas, C. S. "Validity and Usefulness of Student Evaluations of Instructional Quality." *Journal of Educational Psychology,* 1975, *67,* 833–839.

Marsh, H. W., and Roche, L. A. "Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility." *American Psychologist*, 1997, *52*(11), 1187–1197.

Marsh, H. W., and Ware, J. E. "Effects of Expressiveness, Content Coverage, and Incentive on Multidimensional Student Rating Scales: New Interpretations of the Dr. Fox Effect." *Journal of Educational Psychology,* 1982, *74,* 126–134.

McCallum, L. W. "A Meta-Analysis of Course Evaluation Data and Its Use in the Tenure Decision." *Research in Higher Education,* 1984, *21,* 150–158.

McKeachie, W. J. "Student Ratings: The Validity of Use." *American Psychologist,* 1997, *52,* 1218–1225.

Murray, H. G. "Low-Inference Classroom Teaching Behaviors and Student Ratings of College Teaching Effectiveness." *Journal of Educational Psychology,* 1983, *71,* 856–865.

Naftulin, D. H., Ware, J. E., and Donnelly, F. A. "The Doctor Fox Lecture: A Paradigm of Educational Seduction." *Journal of Medical Education,* 1973, *48,* 630–635.

Ory, J. C., Braskamp, L. A., and Pieper, D. M. "The Congruency of Student Evaluative Information Collected by Three Methods." *Journal of Educational Psychology,* 1980, *72,* 181–185.

Overall, J. U., and Marsh, H. W. "Midterm Feedback from Students: Its Relationship to Instructional Improvement and Students' Cognitive and Affective Outcomes." *Journal of Educational Psychology,* 1979, *71,* 856–865.

Overall, J. U., and Marsh, H. W. "Students' Evaluations of Instruction: A Longitudinal Study of Their Stability." *Journal of Educational Psychology,* 1980, *72,* 321–325.

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

Rodin, M., and Rodin, B. "Student Evaluations of Teachers." *Science,* 1972, *177,* 1164–1166.

Scriven, M. "Summative Teacher Evaluation." In J. Milman (ed.), *Handbook of Teacher Evaluation.* Thousand Oaks, Calif.: Sage, 1983.

Seldin, P. "How Colleges Evaluate Professors: 1983 Versus 1993." *AAHE Bulletin,* Oct. 1993a, pp. 6–8, 12.

Seldin, P. "The Use and Abuse of Student Ratings of Professors." *Chronicle of Higher Education,* July 21, 1993b, p. A40.

Williams, W. M., and Ceci, S. J. "How'm I Doing? Problems with Student Ratings of Instructors and Courses." *Change,* 1997, *29*(5), 13–23.

JAMES A. KULIK is director and research scientist for the Office of Evaluations and Examinations at the University of Michigan, Ann Arbor.