Student Ratings

The Validity of Use

Wilbert J. McKeachie
University of Michigan

In this article, the author discusses the other articles in this Current Issues *section and concludes that all of the authors agree that student ratings are valid but that contextual variables such as grading leniency can affect the level of ratings. The authors disagree about the wisdom of applying statistical corrections for such contextual influences. This article argues that the problem lies neither in the ratings nor in the correction but rather in the lack of sophistication of personnel committees who use the ratings. Thus, more attention should be directed toward methods of ensuring more valid use.*

I chuckled with pleasure at some of the thrusts and counterthrusts as I read the preceding articles by Greenwald (1997, this issue), Marsh and Roche (1997, this issue), d'Apollonia and Abrami (1997, this issue), and Greenwald and Gillmore (1997, this issue) in this *Current Issues* section. Each article contains much good sense. My role, presumably, is to give an overview in terms of my experience as a researcher, a teacher, and an evaluator of teaching both for improvement and for personnel decisions. The articles in this section address three main issues: (a) How many dimensions of teaching should student rating forms report to personnel committees? (b) Are student ratings valid measures of teaching effectiveness? and (c) Are student ratings biased by variables other than teaching effectiveness, and if so, can these biases be controlled statistically? I shall briefly address each of these issues. Then I argue that the basic problem is not with the ratings but rather with the lack of sophistication of those using them for personnel purposes. I conclude with some observations and recommendations for research and practice.

How Many Dimensions of Teaching Should Student Rating Forms Report?

The answer to this question depends on what one wants to do with the ratings. Most people interested in improving teaching see the primary purpose of student ratings as providing feedback to teachers that will be helpful for improvement. General overall ratings provide little guidance. Murray (1983, 1997) has shown that specific behavioral items are most likely to result in improvement. Renaud and Murray (1997) have shown that actual behaviors of teachers as coded by observers covary with student ratings of the same behaviors and fall into dimensions corresponding fairly well to those of Marsh (1984). Marsh's demonstration of the validity of these factors is impressive. Grouping items by factors can reduce the "mental dazzle" of a long computer printout of many items and can increase the likelihood of improvement.

But what about reports to committees or administrators making personnel decisions? Such a committee must arrive at a single judgment of overall teaching effectiveness. If one grants that overall ratings of teaching effectiveness are based on a number of factors, should a score representing a weighted summary of the factors be represented (as Marsh and Roche [1997] argue), or should one simply use results of one or more overall ratings of teaching effectiveness (as contended by d'Apollonia and Abrami, 1997)? I would prefer student ratings of attainment of educational goals rather than either of these alternatives. Whatever score or scores are used, I agree with d'Apollonia and Abrami's conclusion: "We recommend that . . . only crude judgments of instructional effectiveness (exceptional, adequate, and unacceptable) [be made on the basis of student ratings]" (p. 1205).

The first reason for a simple three-category classification is that personnel committees do not need to make finer distinctions. The most critical decision requires only two categories—"promote" or "don't promote." Even decisions about merit increases require no more than a few categories, for example, "deserves a merit increase," "deserves an average pay increase," or "needs help to improve."

A second reason for endorsing d'Apollonia and Abrami's (1997) view is that effective teachers come in many shapes and sizes. Scriven (1981) has long argued that no ratings of teaching style (e.g., enthusiasm, organization, warmth) should be used, because teaching effectiveness can be achieved in many ways. Using characteristics that generally have positive correlations with effectiveness penalizes the teacher who is effective despite less than top scores on one or more of the dimensions

I thank Matt Kaplan, Diana Kardia, and James Kulik for their helpful comments on earlier drafts of this article.

Correspondence concerning this article should be addressed to Wilbert J. McKeachie, Department of Psychology, University of Michigan, 525 East University, Ann Arbor, MI 48109-1109. Electronic mail may be sent via Internet to billmck@umich.edu.

 0003-066X/97/$2.00
Vol. 52, No. 11, 1218–1225

usually associated with effectiveness. Judging an individual on the basis of characteristics, Scriven says, is just as unethical as judging an individual on the basis of race or gender.

A third problem with a profile of scores on dimensions is that faculty members and administrators have stereotypes about what good teaching involves. In most meetings to make decisions about promotions or merit salary increases, negative information is likely to be weighted more heavily than positive information. Thus, teachers who do not conform to the stereotype are likely to be judged to be ineffective despite other evidence of effectivenes. My colleagues and I have found evidence of this effect in our studies of the use of student ratings in promotion decisions (Lin, McKeachie, & Tucker, 1984; Salthouse, McKeachie, & Lin, 1978). If personnel committees sensibly use broad categories rather than attempting to interpret decimal-point differences, either a single score or a weighted combination of factor scores will provide comparable results.

Do Student Ratings Provide Valid Data About Teaching Effectiveness?

Evaluation for Improving Teaching

In the articles in this *Current Issues* section, there is little disagreement about the usefulness of student ratings for improvement of teaching (at least when student ratings are used with consultation or when ratings are given on specific behavioral characteristics). There are, however, two problems that detract from the usefulness of ratings for improvement.

The first problem involves students' conceptions of effective teaching. Many students prefer teaching that enables them to listen passively—teaching that organizes the subject matter for them and that prepares them well for tests. Unfortunately, most college teachers are not well trained in test construction. Even teachers who have development of students' thinking as a primary goal give examinations that primarily involve rote memory (McKeachie & Pintrich, 1991).

Cognitive and motivational research, however, points to better retention, thinking, and motivational effects when students are more actively involved in talking, writing, and doing (McKeachie, 1951; Murray & Lan, 1997). Thus, some teachers get high ratings for teaching in less than ideal ways.

The second problem is the negative effect of low ratings on teacher motivation. If a teacher is already anxious, then ratings that confirm the impression that students are bored or dissatisfied are not likely to increase the teacher's motivation and eagerness to enter the classroom and face the students.

A solution for both of these problems is better feedback. Marsh and Roche (1993) demonstrated that feedback targeted to specific problems identified by student ratings results in improvement. Murray and Smith (1989) found that items on specific teaching behaviors resulted in greater improvement than ratings on more general characteristics. In addition, research shows that student ratings are more helpful if they are discussed with a consultant or a peer (Aleamoni, 1978; Cohen, 1980; Marsh & Overall, 1979; McKeachie et al., 1980). Ideally, consultation should be only one feature of an academic culture in which colleagues discuss teaching and both teachers and students develop a sophisticated understanding of what is most helpful for lasting learning.

Evaluation for Promotion

But what about the use of student ratings for personnel decisions? Here again, the authors of the articles in this *Current Issues* section provide reassurance. All of the authors (and I join them) agree that student ratings are the single most valid source of data on teaching effectiveness. In fact, as Marsh and Roche (1997) point out, there is little evidence of the validity of any other sources of data.

However, student ratings are not perfectly correlated with student learning, even in the validity studies carried out in large courses with multiple sections. Many multisection courses use objective tests that assess factual knowledge. In these courses, students' ratings of teaching effectiveness are likely to reflect a relatively unsophisticated conception of effectiveness.

What is effective, however, is more complex. It depends on one's definition of the goals of teaching. If one believes that retention and later use of course concepts are important, mere presentation of the subject and testing for memory of facts is not likely to be effective. If one believes that outcomes such as skills for continued learning and critical thinking, motivation for lifelong learning, and changes in attitudes and values are important, it becomes clear that effective teaching must involve much more student talking, writing, and doing as well as evaluation methods that probe more deeply than most true–false or multiple-choice tests.

I agree with Marsh and Roche's (1997) statement that researchers need to provide validity data that go beyond recall of facts. Both Marsh and I have found student ratings to be valid with respect to other criteria, including motivational, attitudinal, and other goals of education (Marsh, 1984; McKeachie, Guetzkow, & Kelly, 1954; McKeachie, Lin, & Mann, 1971; McKeachie & Solomon, 1958).

The good news is that student ratings correlate positively with these indexes of teachers' effectiveness. The bad news is that teachers are not equally effective for all goals and all students. Cross (1958) found in one multisection study that there was a negative correlation between teachers' effectiveness as measured by the multiple-choice portion of the final examination and effectiveness as measured by the essay portion of the final examination. Hoyt and Cashin (1977) found that teaching behaviors associated with learning factual knowledge were different from those that help students develop problem-

solving skills or self-understanding. Thus, a personnel committee needs to consider the relative importance of different educational goals when assessing teaching.

In addition to the need to look at other outcomes, researchers need to be aware of two additional problems with multisection studies. The first problem is that multisection courses are primarily first- and second-year courses; student ratings in these courses may have lower validity coefficients than in more advanced courses in which students have broader experience (and perhaps greater educational sophistication) as a basis for their ratings.

The second problem is that the achievement measure is common to all sections. Thus, what it assesses with respect to teaching is how well the teacher has prepared students for the test; it does not assess learning that goes beyond the test. And the test almost necessarily must be based on common material in the textbook. A classic study by Parsons (1957) found that students who simply studied the textbook without any classroom instruction did better on the final course examination than did those who had conventional classroom instruction that went beyond the textbook.

Isaacson, McKeachie, and Milholland (1963) found that the teaching assistants who were most effective had been rated by their peers as having broad cultural interests and knowledge. Good teachers often go well beyond the textbook. To get a valid measure of real teaching effectiveness, researchers need to measure not only what is taught in common but also educational gains that go beyond the minimum measured by a common examination. Students' papers, journals, and measures of motivation and attitude or other outcomes are needed.

Not only do good teachers go beyond the textbook, but their influence goes well beyond the geographical confines of the classroom. Most student rating forms and most faculty members' evaluations of teaching effectiveness focus almost completely on conventional classroom teaching. Clearly, much—very likely most—student learning occurs outside the classroom. Researchers need to get data on teachers' out-of-class contributions to education (d'Apollonia and Abrami's, 1997, "teacher as manager"). Student rating forms need to cue students to consider these aspects of teaching in their ratings.

Are Student Ratings Biased by Other Variables?

Greenwald and Gillmore (1997) are concerned about at least two sources of bias—class size and grading leniency. The concern about class size seems to me to be valid only if a personnel committee makes the mistake of using ratings to compare teachers rather than as a measure of teaching effectiveness. There is ample evidence that most teachers teach better in small classes. Teachers of small classes require more papers, encourage more discussion, and are more likely to use essay questions on examinations—all of which are likely to contribute to student learning and thinking. Thus, on average, small classes should be rated higher than large classes.[1]

Grading bias, however, is a more serious problem. I have little doubt that giving higher grades can raise ratings if one can convince students that they have learned more than is typical. But students are not so likely to be positively affected if an ineffective teacher seems to be trying to buy good ratings with easy grades. In fact, the attempt may boomerang. A former faculty member whose grades were the highest in my department received the lowest student ratings; Abrami, Dickens, Perry, and Leventhal (1980) presented more systematic evidence of the negative effect of giving undeserved higher grades.

The effect of easy grading may well depend on the institution. Clark and Trow (1966) demonstrated that colleges and universities differ in their dominant cultures: some emphasizing academic values, others emphasizing social and collegiate values. If students have primarily chosen a college to have a good time, easy teachers may be more highly appreciated than in institutions with stronger academic cultures.

Whether or not student ratings are positively affected by grading leniency, the effect on a promotion committee's judgments is likely to be much more negative if the committee perceives the grading pattern to be higher than normal. Even when Sullivan (1974) had convincing evidence that students in his programmed learning class achieved more than students in conventional classes, he encountered fierce hostility from his colleagues about giving higher grades. Faculty members and administrators are concerned about possible grade inflation. Good student ratings accompanied by a higher than normal grade distribution are likely to be a ticket to termination before tenure.[2]

Can Something Be Done to Prevent the Success of Those Who Attempt to Buy Higher Ratings From Students With High Grades?

Greenwald and Gillmore (1997) suggest that only the grading-bias hypothesis can explain four patterns in correlational data and thus justify the use of a statistical correction. Unfortunately, their argument that only the grading-bias hypothesis can account for their four findings seems to me to be flawed. I examine each in turn.

[1] Greenwald and Gillmore (1997) are concerned that even though a teacher who teaches a small class may be more effective, this makes for an unfair advantage when that teacher is compared with a teacher of a large class. I argue that the mistake is in making such comparisons rather than in a bias in the student ratings.

[2] Greenwald (1997) suggests that most personnel committees are not aware of differences in grading standards. It may be that this varies among institutions. Certainly it has come up a number of times in my experience as a department chair and a committee member. I asked one of the senior members of the faculty at the University of Michigan who not only had experience on committees at the University of Michigan but also had chaired a department at another major university about his experience, and he said that grading leniency did come up frequently when discussing a faculty member's teaching.

Positive Grades–Ratings Relationships Within Classes

Here, the assumption is that the teacher is equally effective for all students within a class. In fact, there are numerous studies that have shown attribute–treatment interactions. With specific reference to the within-class correlations, Remmers (1928) suggested, and Elliott (1950) showed, that within-class correlations between grades and student ratings are a function of the level at which the instructor pitches the class. If the instructor teaches primarily to the better students (as many teachers do), then these students achieve more than expected and rate the instructor more highly than do other students, resulting in a positive correlation. By contrast, in a class where the teacher helps poorer students achieve more than predicted, these students give the instructor higher ratings, resulting in a negative correlation between ratings and grades. Greenwald and Gillmore's (1997) results support the common impression that many teachers teach to the better students; in fact, it has not been long since first-year courses (particularly in the sciences) were designed to weed out students who did not belong in those disciplines.

Stronger Grades–Ratings Relationships With Relative, Rather Than Absolute, Measures of Expected Grade

If students feel that they are learning more in a particular class than they are in other classes, it should not be surprising that they will rate teaching effectiveness higher in the former class. Why, then, is not the actual grade expected as highly correlated with ratings as the relative grade? This is likely to be true if teachers strike a chord with some students whose performance in other classes is average or below average. These students will rate the teacher highly and expect their grades to be higher than normal, but the actual expected grades will still not be As. Again, the teaching-effectiveness hypothesis is not disconfirmed by this result.

Grade-Related Halo Effect in Judging Course Characteristics

I admire Greenwald and Gillmore's (1997) ingenuity in thinking of this analysis. Nevertheless, their argument seems to me to be irrelevant to the validity of between-course ratings. As Greenwald and Gillmore point out, students tend to blame the instructor if they fail to learn; thus, it is not surprising that they find fault with many characteristics of the teacher. Those who are having the most difficulty are most likely to blame the situation, resulting in a negative halo. Nevertheless, it may be stretching my attribute–treatment interaction hypothesis too far to explain the halo within, but not between, classes.

Because the focus of the ratings is on overall teaching effectiveness, one should not be surprised to find a halo effect. The appropriate question is as follows: Does the halo effect invalidate students' overall ratings of teaching effectiveness? It probably does to the degree that concern for students' learning and other positive teacher characteristics are overweighted by students in their overall judgment. Thus, those students (frequently the less able) who feel that the teacher does not care about their learning develop a negative halo, whereas those who feel that the teacher cares about them develop a positive halo. However, this does not bear on the validity of the overall rating. In fact, as d'Apollonia and Abrami (1997) point out, the halo effect may increase validity.

Negative Grades–Workload Relationship Between Classes

The negative relationship between grades and workload is not directly relevant to the issue of grading leniency. Part of the relationship is probably due to aggregating data across departments. Science departments tend to give lower grades than humanities and social science departments and are perceived by students as requiring more work (Cashin & Sixbury, 1993; Centra, 1993). Within most departments, there are also ineffective teachers who, feeling alienated from their students, require more work and then blame their students for not meeting the teachers' standards.

Greenwald and Gillmore (1997) assume that hours worked should relate to learning and grades. They probably do. Unfortunately, the relationship is not a simple one. In general, one would expect that students who are having difficulty will spend more time studying than will those who have better background knowledge. This is likely to result in better learning for the less able students but is not likely to result in the kind of positive relationship between workload and grades that Greenwald and Gillmore expected.

Although the workload–grades relationship does not involve student ratings, Greenwald and Gillmore (1997) apparently draw the implication that low-workload courses will be given high ratings. In interviews of students, I have found that often the workload is heavy because the teacher has been ineffective—assignments are unclear, lectures are disorganized, and tests require memorization of definitions and a myriad of specific facts. Thus, Greenwald and Gillmore need to differentiate between hours spent compensating for poor instruction and work that is constructive in promoting learning and increasing motivation. Greenwald and Gillmore could distinguish between these two kinds of "work," I believe, by looking at ratings on such items as "I increased my interest in this field." Again, the teaching-effectiveness hypothesis is not disconfirmed.

Conclusion

Both the grading-bias and teaching-effectiveness hypotheses can account for Greenwald and Gillmore's (1997) findings. Nonetheless, I agree with them that grading leniency can sometimes affect ratings. If the correlation between mean grades and ratings were due only to intentional efforts to get higher ratings, a statistical correction

would be appropriate. However, there are at least two kinds of cases in which such a correction would be inappropriate—the excellent teacher whose students' achievement merits higher grades and the poorer teacher whose grades are unjustifiably low. For most teachers, the correction would make little difference. Just as in controlling students' cheating, evaluators should focus on preventive measures rather than implementing measures that will punish effective teachers as well as those who cheat.

Preventing Cheating

What can be done to reduce the sort of desperation that leads to cheating? Clearly, the most desirable measure would be to increase teachers' competence so that student ratings are validly positive, thus reducing the temptation to cheat. This implies strategies such as better preparation for college teaching in graduate school, better orientation and training during the first years of teaching, and collecting student ratings early in the term and discussing them with a consultant or fellow teacher.

The temptation to cheat also may be affected by faculty members' confidence that the judgments will be fair. To make sure that contextual variables influencing ratings are taken into account, personnel committees should consider teachers' own statements about the goals they were trying to achieve, how they went about achieving them, and the contextual conditions that might have influenced success.[3] As Cashin (1995) suggested in his review of the research on correlations between expected grades and ratings, the best method of control is to review graded course materials to judge whether the standards are appropriate.

One would like the committee's judgments to be based on valid evidence. Student ratings are valid, but all of the authors in this *Current Issues* section agree that they should be supplemented with other evidence. Yet, as Marsh and Roche (1997) point out, there is little research on the validity of other sources of evidence. Clearly, such research is needed.

The Validity of Use of Ratings in Personnel Decisions

The authors of the articles in this *Current Issues* section agree that student ratings are the most valid and practical source of data on teaching effectiveness. But, as I noted earlier, these data must then be interpreted by faculty or administrators who must make decisions about promotions and merit pay increases.

I contend that the specific questions used, the use of global versus factor scores, the possible biasing variables, and so forth are relatively minor problems. The major validity problem is in the use of the ratings by personnel committees and administrators (Franklin & Theall, 1989).

No matter how valid the evidence provided by students may be, it is almost certainly more valid than many personnel committees give it credit for being. I have participated in more than 1,000 reviews of faculty members for promotions or merit pay increases. In my opinion, many committees seem to make sensible use of student rating results, but all too often, I have heard student ratings dismissed with such phrases as "He's not a good researcher—obviously he can't be an excellent teacher," "You can't expect students to know which teachers were good until they've been out of college a few years," or "All students want are some jokes and an easy grade." Whatever the reason, student ratings of teaching are often not given heavy weight in promotion decisions.

Although I believe that a statistical adjustment of ratings, such as Greenwald and Gillmore (1997) suggest, may result in lower, rather than higher, validity, it may increase the credibility of the ratings. If it thus contributes to better weighting of ratings in personnel decisions, I'm for it.

Almost as bad as dismissal of student ratings, however, is the opposite problem—attempting to compare teachers with one another by using numerical means or medians. Comparisons of ratings in different classes are dubious not only because of between-classes differences in the students but also because of differences in goals, teaching methods, content, and a myriad of other variables. Moreover, as I suggested earlier, comparisons are not needed for personnel decisions. To the degree that student ratings enter into such decisions, faculty members can be reliably allocated to three or four categories (as d'Apollonia and Abrami [1997] suggest) by simply looking at the distribution of student ratings: How many students rated the teacher as very good or excellent? How many students were dissatisfied?

What Can Be Done to Improve the Validity of the Use of Student Ratings?

Presumably the result educators would like to achieve is appropriate recognition of teaching in personnel decisions, and until those making the decisions become more sophisticated, the nature of the instrument and possible biases are not likely to make significant differences. Research at the University of Michigan on the use of ratings in personnel decisions has used simulated dossiers rated by individual members of committees determining promotions (Lin et al., 1984; Salthouse et al., 1978). But as far as I know, there has been no research on the actual decision-making processes in the committees. It would be difficult, but perhaps not impossible, to obtain permission to carry out observational studies of actual meetings of such committees. If this proves to be impossible, it should be possible to carry out research using simulated meetings in which experimental variations could be tested.

[3] There are probably some classrooms where no one could get top ratings! I once taught a spring class in which the room was unbearably hot if the windows were closed and unbearably noisy from the jackhammers nearby when the windows were open.

If one were to carry out a program of such research with some design that enabled one to discriminate more valid from less valid outcomes, I would not be surprised to find that one would emerge with results similar to those in studies of medical diagnosis and mortality predictions. Either a computer program or a pooled judgment of physicians tends to be superior to predictions of individual physicians. However, the combination of the computer program and the physicians is better yet (Yates, 1994). Thus, I can envision a time when promotion decisions are made by using a weighted combination of Marsh and Roche's (1997) factors along with the pooled judgment of well-trained committee members.

That time is not near, and in the meantime, researchers need to improve the quality of the data presented. The research of Greenwald and Gillmore (1997), Marsh and Roche (1997), d'Apollonia and Abrami (1997), and others has contributed greatly to understanding student ratings of instruction, but in addition to research on student ratings, research is needed on ways of teaching students to be more sophisticated evaluators as well as ways that the experience of filling out the rating form can become more educational for students. For example, qualitative research on what goes on in students' minds when they are filling out evaluations would provide a better idea of whether they are analyzing their own learning or are simply discharging a boring chore. What kinds of items, what kinds of structure for ratings, and what balance of ratings and open-ended questions would stimulate more thought?

Student rating forms have mostly been developed using the approach of "dust bowl empiricism," that is, get a number of items about teaching and see what works. During the 1950s, I tried to collect every student rating form then in use in the United States, and Isaacson et al. (1963) then factor analyzed all of the items I had gathered. I still believe this was a useful approach, but in the 1990s, there are much better theories of cognition and motivation, and student rating forms should now better reflect those theories. Although I have stressed that validity of use is the key issue, researchers should also be looking, as Marsh and Roche (1997) have, at construct validity with respect to theories of teaching.

Even this, however, probably will not be sufficient to handle all the modes of teaching. The increasing use of technology, virtual universities, studio teaching, clinical teaching, cooperative learning, and service learning represents important aspects of education, and we very likely need a variety of forms and items to accommodate such differences.

For summative evaluations by personnel committees, I like the method developed by Hoyt, Owens, Cashin, and others at the Center for Faculty Evaluation and Development at Kansas State University—IDEA (Instructional Development and Effectiveness Assessment). The IDEA form asks students to rate their progress on each of 10 instructional goals—a method that not only provides information that goes beyond the teacher's conformity to a naive stereotype of good teaching but also is educational in broadening students' conceptions of what the aims of education are. Students may not always be able to accurately assess their own progress, but if asked, they do know whether a course mainly required memorization or thinking, and they should know whether a course increased their interest in further learning in that subject-matter area. Use of such items about goals appropriately leaves to the personnel committee the judgment as to which goals are most important for a particular course in the context of the overall objectives of the department and the university. Student ratings of their attainment of educational objectives not only provide better data for personnel committees but also stimulate both students and teachers to think about their objectives—something that is educational in itself.

Researchers also need to study what teachers can do to help students become more sophisticated raters. As I have pointed out, many faculty members and students have rather limited notions of the goals of education and of what is conducive to learning that will last and be used. Thus, faculty need to be educated, and then they need to be encouraged to explain to their students why the requirements they make and the procedures they use are likely to contribute to better learning.

Most of all, research is needed on how to train members of personnel committees to be better evaluators, and research is needed on ways of communicating the results of student evaluations to improve the quality of their use. I was pleased to note that at the 1997 meeting of the American Educational Research Association, some papers were beginning to address these problems of use. For example, Jennifer Franklin (personal communication, March 26, 1997) reported that she is now presenting the ratings and the confidence levels in graphic form to overcome the problem of misinterpretation of numerical means and norms. Katherine Ryan (1997) studied faculty members' views of different reporting approaches and found that faculty members would prefer a standards-based approach rather than norm-referenced reports. As d'Apollonia and Abrami (1997) note, Abrami and I have argued (McKeachie, 1996) that the use of norms not only leads to comparisons that are invalid but also is damaging to the motivation of the 50% of faculty members who find that they are below average. Moreover, presentation of numerical means or medians (often to two decimal places) leads to making decisions based on small numerical differences—differences that are unlikely to distinguish between competent and incompetent teachers.

With respect to my plea for training members of personnel committees, Villaescusa, Franklin, and Aleamoni (1997) reported that a workshop for faculty and administrators improved knowledge and opinions about student ratings. Unfortunately, Ryan (1997) found that most faculty members would not be interested in attending such a workshop. We need research on methods, in addition to workshops, that can help increase the valid use of ratings in personnel decisions. Could such commit-

tees be persuaded to accept consultants who would assist them in interpreting student ratings but not take part in the actual decision making?

There now is ample evidence of ways in which teaching can be improved. The problem is how to get research findings into use. Therefore, research and theory is needed not only on the nature and measurement of good teaching but also on the problems of getting theory into use—use in training teachers; use in personnel decisions; and use in methods of collecting data about teaching, such as portfolios, classroom observations, assessments of syllabi, tests, and course materials, and student rating forms.

Conclusion

As Herb Simon (1997) recently said, "Learning is ultimately a human activity, regardless of the technology used." Students will continue to be those most affected by teaching. Therefore, student ratings will continue to be useful.

I end by considering once again Greenwald's (1997) experience that initiated this set of articles. He was surprised that he received markedly lower ratings in one course than in another course that he had taught in the same way. Had I been consulting with him about the ratings, I would have said something like this:

Tony, classes differ. Effective teaching is not just a matter of finding a method that works well and using it consistently. Rather, teaching is an interactive process between the students and the teacher. Good teaching involves building bridges between what is in your head and what is in the students' heads. What works for one student or for one class may not work for others. Next time, get some ratings early in the term, and if things are not going well, let's talk about varying your strategies.

Fortunately, I was not his consultant, and the result was the series of research studies he and Gillmore (1997) reported as well as the initiation of this group of articles.

REFERENCES

Abrami, P. C., Dickens, W. J., Perry, R. P., & Leventhal, L. (1980). Do teacher standards for assigning grades affect student evaluations of instruction? *Journal of Educational Psychology, 72,* 107–118.

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

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited* (IDEA Paper No. 32). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E., & Sixbury, G. R. (1993). *Comparative data by academic field* (IDEA Tech. Rep. No. 8). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Centra, J. A. (1993). *Reflective faculty evaluation.* San Francisco: Jossey-Bass.

Clark, B., & Trow, M. (1966). The organizational context. In T. M. Newcomb & E. K. Wilson (Eds.), *College peer groups: Problems and prospects for research* (pp. 17–70). Chicago: Aldine.

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.

Cross, D. (1958). *An investigation of the relationships between students' expressions of satisfaction with certain aspects of the college classroom situation and their achievement on the final examination.* Unpublished honors thesis, University of Michigan.

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

Elliott, D. H. (1950). Characteristics and relationships of various criteria of colleges and university teaching. *Purdue University Studies in Higher Education, 70,* 5–61.

Franklin, J., & Theall, M. (1989, April). *Who read ratings: Knowledge, attitude and practice of users of student ratings of instruction.* Paper presented at the 70th annual meeting of the American Educational Research Association, San Francisco.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52,* 1182–1186.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,* 1209–1217.

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

Isaacson, R. L., McKeachie, W. J., & Milholland, J. M. (1963). Correlation of teacher personality variables and student ratings. *Journal of Educational Psychology, 54,* 110–117.

Lin, Y.-G., McKeachie, W. J., & Tucker, D. G. (1984). The use of student ratings in promotion decisions. *Journal of Higher Education, 55,* 583–589.

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., & 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 student evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 30,* 217–251.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52,* 1187–1197.

McKeachie, W. J. (1951). Anxiety in the college classroom. *Journal of Educational Research, 45,* 135–160.

McKeachie, W. J. (1996). Do we need norms of student ratings to evaluate faculty? *Instructional Evaluation and Faculty Development, 14,* 14–17.

McKeachie, W. J., Guetzkow, H., & Kelly, E. L. (1954). An experimental comparison of recitation, discussion and tutorial methods in college teaching. *Journal of Educational Psychology, 45,* 224–232.

McKeachie, W. J., Lin, Y.-G., Daugherty, M., Moffett, M., Neigler, C., Nork, J., Walz, M., & Baldwin, R. (1980). Using student ratings and consultation to improve instruction. *British Journal of Educational Psychology, 50,* 168–174.

McKeachie, W. J., Lin, Y.-G., & Mann, W. (1971). Student ratings of teaching effectiveness: Validity studies. *American Educational Research Journal, 8,* 435–445.

McKeachie, W. J., & Pintrich, P. (1991). Program on classroom teaching and learning strategies. In J. S. Stark & W. J. McKeachie (Eds.), *Final report: National Center for Research to Improve Postsecondary Teaching and Learning* (pp. 41–59). Ann Arbor: University of Michigan, School of Education.

McKeachie, W. J., & Solomon, D. (1958). Student ratings of instructors: A validity study. *Journal of Educational Research, 51,* 379–382.

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, March). *Classroom teaching behaviors and student instructional ratings: How do good teachers teach?* McKeachie Award address presented at the 78th annual meeting of the American Educational Research Association, Chicago.

Murray, H. G., & Lan, M. (1997). The relationship between active participation and student learning. *STLHE/SAPES, 20,* 7–10.

Murray, H. G., & Smith, T. A. (1989, April). *Effects of midterm behavioral feedback on end-of-term ratings of instructor effectiveness.* Paper presented at the 70th annual meeting of the American Educational Research Association, San Francisco.

Parsons, T. S. (1957). A comparison of learning by kinescope, correspondence study, and customary classroom procedures. *Journal of Educational Psychology, 48,* 27–40.

Remmers, H. H. (1928). The relationships between students' marks and students' attitudes toward instructors. *School and Society, 28,* 759–760.

Renaud, R. D., & Murray, H. G. (1997, March). *Factorial validity of student ratings of intruction.* Paper presented at the 78th annual meeting of the American Educational Research Association, Chicago.

Ryan, K. E. (1997, March). *Making student ratings comprehensible to faculty: A review of alternative reporting approaches.* Paper presented at the 78th annual meeting of the American Educational Research Association, Chicago.

Salthouse, T. A., McKeachie, W. J., & Lin, Y.-G. (1978). An experimental investigation of factors affecting university promotion decisions. *Journal of Higher Education, 49,* 177–183.

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 244–271). Beverly Hills, CA: Sage.

Simon, H. (1997, March). *The future of education in the 21st century.* Paper presented at the Celebration of the 50th Anniversary of the Founding of the American Institutes of Research, Washington, DC.

Sullivan, A. M. (1974). Psychology and teaching. *Canadian Journal of Behavioral Science, 6,* 1–29.

Villaescusa, T., Franklin, J., & Aleamoni, L. (1997, March). *Improving the interpretation and use of student ratings: A training approach.* Paper presented at the 78th annual meeting of the American Educational Research Association, Chicago.

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 381–409). New York: Wiley.