12

Student Ratings: What More Than 80 Years of Research Tell Us

Student ratings are one of the most common features of faculty evaluation systems. More than 85% of all faculty evaluation systems make regular and routine use of student ratings (Seldin, 1993). Despite the continued recommendations in the literature that faculty evaluation systems use multiple sources of data, many working faculty evaluation systems continue to use student ratings of instructor and instructions as their only component. It is not surprising, therefore, that student ratings have tended to become synonymous with faculty evaluation. And, because the evaluation of faculty performance, especially for the purpose of making personnel decisions, is not always the most popular and well-received administrative activity, faculty and administrators continue to ask questions concerning the factors that may, or may not, influence student ratings of instructors and instruction.

The literature concerning student ratings is immense. Far more than 2,000 articles have been written on this one aspect of faculty evaluation alone (Cashin, 1999). Although the great body of research on student ratings has taken place in the last 30 years, studies on student ratings go back more than 80 years (Guthrie, 1954, cites a study on student ratings dating all the way back to 1924). Even today the design, development, use, and interpretation of student ratings continues to be one of the most heavily researched topics in the general area of faculty evaluation. The results of all this research have led to the fairly firm conclusion, at least among the educational research community, that properly constructed, appropriately administered, and correctly interpreted student ratings can be valid and reliable measures indicating the quality of teaching (Aleamoni, 1999).

With all the evidence for the validity and reliability of student ratings in the literature, the basic belief in higher education that student ratings are not valid persists. Before providing an overview and summary of what the large body of research on student ratings says to us, it is useful to examine why this belief persists.

What Is Validity?

Decades of research have demonstrated that tools used in faculty evaluation systems, especially student rating forms, can be designed to be valid and reliable. So why does the belief that student ratings are not valid still reverberate throughout higher education despite so much research? My experience with hundreds of colleges and universities facing the task of revising or building their faculty evaluation systems has led me to conclude that there are three main reasons:

1) Many of the tools used in faculty evaluation systems are homemade and are thus of dubious validity and reliability.

2) Most academic administrators are not conversant with the finer points of psychometrics.

3) Higher education has yet to establish a universally accepted definition of the characteristics and skills necessary for teaching excellence.

These three conditions lead to a cascading set of circumstances that profoundly affect the general professoriate's perception, and willingness to accept, the possibility that

student rating data is, or can be, valid. A brief examination of these conditions reveals the underlying problem.

Homemade Faculty Evaluation Tools

Faculty evaluation tools have, by and large, been dominated by the use of some form of student rating form. Student rating forms have been, and continue today, to be the primary common element of all faculty evaluation systems. In many cases student ratings constitute the only systematically gathered data used in a faculty evaluation system. Unfortunately, the great bulk of student rating forms in use across higher education in America today are homemade. That is, they have been constructed by committees comprised of various combinations of faculty, academic administrators, and students. Owing to the lack of psychometric expertise or rigor in constructing these forms, they are of dubious (and often undetermined) validity and reliability. This common situation has resulted in a rich and voluminous storehouse of anecdotes and stories leading to a number of myths, including the ever-popular myths that student ratings are just a popularity contest and that faculty can buy good ratings by giving easy grades. Unfortunately, these and many other such myths likely have a basis in fact since they may possibly be true for poorly constructed forms. The anecdotes that have produced these and other myths about faculty evaluation tools are so common, so voluminous, and so widespread throughout the culture of higher education that they have taken on an aura of common knowledge. Thus, in the minds of the professoriate and academic administrators, this common knowledge is so pervasive that it far overshadows the truth concerning student ratings and other faculty evaluation tools that are buried in the pages of psychometric journals.

What Is Psychometrics?

It is unfortunate but true that the majority of academic administrators are unfamiliar with the finer points of psychometrics. Deans and vice presidents for academic affairs are usually the ones faced with the task of gathering some form of evaluative information concerning their faculty's performance for the purpose of making promotion, tenure, continuation, or similar personnel decisions. Of the many hundreds of academic administrators with whom I have had the occasion to interact on the issue of faculty evaluation, I could number on one hand the ones that have been sufficiently conversant with psychometrics to understand the subtle differences between, say, content validity and construct validity. These individuals (biologists, musicians, historians, physicists, physicians, etc.) not only are generally unfamiliar with the finer points of psychometrics but rarely, if ever, read articles in such publications as *Journal of Educational Psychology* (American Psychological Association) or *Review of Educational Research* (American Educational Research Association).

From the perspective of many academic administrators responsible for making significant decisions concerning the design, structure, and format of faculty evaluation systems (especially those administrators that come from the so-called "hard sciences"), the entire field of research that speaks to the issues of psychological measurement is, at best, a little-known area and, at worst, considered a phony or illegitimate area of study on par with research on ghosts and leprechauns. Thus, when the issue of the validity of student ratings is raised, the definition or conception of validity used is much more likely to be that of colloquial usage rather than technical precision. It is useful to look at a dictionary definition of *valid* and the technical, psychometric definition of *validity:*

Main Entry: val·id [Dictionary definition]

1 : having legal efficacy or force; *especially* : executed with the proper legal authority and formalities <a *valid* contract>

2 a : well-grounded or justifiable : being at once relevant and meaningful <a *valid* theory> b : logically correct <a *valid* argument> <*valid* inference>

3 : appropriate to the end in view : EFFECTIVE <every craft has its own *valid* methods>

VALID implies being supported by objective truth or generally accepted authority <a *valid* reason for being absent> <a *valid* marriage> (Miriam-Webster OnLine, 2005–2006)

Validity. [Psychometric technical definition] The effectiveness of the test in representing, describing, or predicting the attribute that the user is interested in. *Content validity* refers to the faithfulness with which the test represents or reproduces an area of knowledge. *Construct validity* refers to the accuracy with which the test describes an individual in terms of some psychological trait or construct. *Criterion-related validity,* or *predictive validity,* refers to the accuracy with which the test scores make it possible to predict some criterion variable of educational, job, or life performance. (Thorndike & Hagen, 1969, pp. 163–177).

Looking at the differences between these definitions, and realizing that the majority of faculty and academic administrators use the dictionary definition rather than the psychometric one, it is easy to see why the belief that student ratings are not valid persists. There is no body of anecdotes and stories comparable to the myths supporting the proposition that faculty evaluation data (especially student ratings) are "well-grounded or justifiable, being at once relevant and meaningful, logically correct, or supported by objective truth or generally accepted authority."

What Does the Research Tell Us?

Having acknowledged the cultural aspects of the higher education community that lead to the constant doubting of the utility of student ratings, we can now explore some of the more common questions about student ratings and examine what the research tells us.

In interpreting the research findings relative to the most common questions concerning student ratings, it is important to note that the conclusions drawn are based on the assumed use of professionally developed, valid, and reliable student rating forms. What the literature demonstrates to be a reliable finding concerning how student ratings work may, in fact, not be true for student rating forms that have been homemade by student, faculty, or administrative groups. Such forms have generally not undergone the rigorous psychometric and statistical procedures required to construct a valid and reliable rating. Most of the answers provided to the following common questions concerning student ratings are derived from several reviews of the literature, especially Lawrence M. Aleamoni's (1999) excellent review of the literature on student rating research. The questions addressed are:

- Aren't student ratings just a popularity contest?
- Aren't student rating forms just plain unreliable and invalid?
- Aren't students too immature, inexperienced, and capricious to make any consistent judgments about the instructor and instruction?
- Isn't it true that I can buy good student ratings just by giving easy grades?
- Isn't it generally easier to get good ratings in higher level courses?
- Isn't it true that students who are required to take a course tend to rate the course more harshly than those taking it as an elective?
- Isn't there a gender bias in student ratings? Don't female faculty tend to get lower ratings than male faculty?
- Isn't it more difficult for math and science faculty to get good ratings?
- Isn't it true that the only faculty who are really qualified to teach or evaluate their peers' teaching are those who are actively involved in conducting research in their field?
- Don't students have to be away from the course, and possibly the college, for several years before they are able to make accurate judgments about the instructor and instruction?
- Isn't it true that the size of the class affects student ratings?
- Does the time of the day the course is taught affect student ratings?
- Do majors in a course rate it differently than nonmajors?
- Does the rank of the instructor (instructor, assistant professor, associate professor, or professor) affect student ratings?
- Why bother with 20 or 30 questions on a student rating form? Can't we just use single general items as accurate measures of instructional effectiveness?
- What good are student ratings in efforts to improve instruction?

Aren't student ratings just a popularity contest? The answer to this first most vexing of questions is "no"—if you are using a student rating form that has been constructed using professional psychometric procedures and has demonstrated reliability and validity. Well-designed student rating forms (i.e., those designed according to professional psychometric standards) carefully measure many different aspects of faculty performance. While these may include the approachability or friendliness of a faculty member, which might be interpreted as a popularity issue, well-designed forms also measure many other factors that are unrelated to the popularity issue. Thus, the complete student rating is *not* simply a measure of popularity, but a measure of various instructional design and delivery skills and characteristics. Many studies show that, given a well-designed form, students can be reliable judges of instructional effectiveness.

There is one qualification, however: Student rating forms that have *not* been constructed according to professional psychometric standards may be unreliable and thus

able to be influenced by such factors as popularity, temperature of the classroom, instructor gender, or anything else. Unfortunately, many institutions do use student rating forms that have not been constructed and validated using professional psychometric standards. Without rigorous reliability and validity data on such forms, it is impossible to tell for certain what influences the final student rating. (See: Aleamoni, 1999; Aleamoni & Spencer, 1973; Aleamoni, 1976; Arreola, 1983; Feldman, 1989; Tang, 1997; Beatty & Zahn, 1990; Benz & Blatt, 1995, Costin, Greenough, & Menges, 1971; Dukes & Victoria, 1989; Frey, 1978; Grush & Costin, 1975; Johannessen & Associates, 1997; Krehbiel & Associates, 1997; Macdonald, 1987; Marlin, 1987; Marsh & Bailey, 1993; Perry, Abrami, & Leventhal, 1979; Rodabaugh & Kravitz, 1994; Shepherd & Trank, 1989; Tollefson, Chen, & Kleinsasser, 1989; Ware & Williams, 1977; Waters, Kemp, & Pucci, 1988.)

Aren't student rating forms just plain unreliable and invalid? The answer to this question is, surprisingly, "yes and no." Most student rating forms in use today suffer from unreliability and invalidity problems. As noted in the earlier discussion concerning the issue of validity, the great preponderance of student rating forms used by colleges and universities have been developed by committees comprised of students, faculty, and/or administrative groups who have not followed the rigorous psychometric and statistical procedures required to produce a valid and reliable student rating form. Well-developed instruments have been shown to be reliable and valid. Costin et al. (1971) and Marsh (1984) reported reliabilities for such forms to be about 0.90. Aleamoni (1978a) reported reliabilities ranging from 0.81 to 0.94 for items and from 0.88 to 0.98 for subscales of the Course/Instructor Evaluation Questionnaire (CIEQ; see www.cieq.com). It should be noted, however, that wherever student rating forms are not carefully constructed with the aid of professionals, as in the case of most student- and faculty-generated forms (Everly & Aleamoni, 1972), the reliabilities may be so low that they completely negate the evaluation effect and its results.

Validity, in the psychometric sense, is much more difficult to assess than reliability. Most student rating forms have been validated by the judgment of experts that the items and subscales measure important aspects of instruction (Costin et. al., 1971). These subjectively determined dimensions of instructional setting and process have also been validated using statistical tools, such as factor analysis (Aleamoni & Hexner, 1980; Burdsal & Bardo, 1986; Marsh, 1984). Further evidence of validity comes from studies in which student ratings are correlated with other indicators of teacher competence, such as peer (colleague) ratings, expert judges' ratings, graduating seniors' and alumni ratings, and student learning. The 14 studies cited by Aleamoni and Hexner (1980) in which student ratings were compared to colleague ratings, expert judges' ratings, graduating seniors' and alumni ratings, and student learning measures all indicated the existence of moderate to high positive correlations, which can be considered as providing additional evidence of validity. The one or two studies that have found a negative relationship between student achievement and instructor rating have been found to be methodologically flawed by several researchers (Centra, 1973b; Frey, 1973; Gessner, 1973; Menges, 1973).

Aren't students too immature, inexperienced, and capricious to make any consistent judgments about the instructor and instruction? The answer to this question is "no." Evidence dating back to 1924, according to Guthrie (1954), indicates that this commonly held belief is not true. The stability of student ratings from one year to the next results in substantial correlations in the range of 0.87 to 0.89. More recent literature on the subject cited by Costin, Greenough, and Menges (1971) and studies by Gillmore (1973) and Hogan (1973) indicated that the correlation between student ratings of the same instructors and courses ranged from 0.70 to 0.87.

Isn't it true that I can buy good student ratings just by giving easy grades? There has been more research conducted on this one question than almost any other in the field of student ratings and faculty evaluation. Nearly 500 studies have been conducted in trying to answer this question. The reason for the numerous studies is not that the question is so difficult to answer—it's that faculty generally don't like the answer the literature provides. The answer is that there is *no* consistent correlation between the grades a faculty member gives and the ratings he or she receives from a well-designed student rating form. No clear positive correlation between grades and student ratings has emerged in the literature. The bulk of the correlations appearing in the literature have been quite weak ranging from -.18 to + .18 (i.e., average r = 0.0).

Considerable controversy has centered around the relationship between student ratings and actual or expected course grades. The general feeling is that students tend to rate courses and instructors more highly when they expect to, or actually receive, good grades. Correlational studies have reported widely inconsistent grade-rating relationships. Twenty-two studies have reported zero relationships (Aleamoni & Hexner 1980), and 28 studies have reported significant positive relationships (Aleamoni & Hexner, 1980). In most instances however, these relationships were relatively weak, as indicated by the fact that the median correlation was approximately 0.14, with the

mean and standard deviation being 0.18 and 0.16, respectively. The clear outcome from the studies on this issue is that, at best, the relationship between grades and ratings is extremely weak, with the average correlation across all studies being 0.0. Clearly, the idea that student ratings are highly correlated with grades is not supported by the literature (Frey, 1973; Gessner, 1973; Sullivan & Skanes, 1974). Again, however, it must be noted that we can't be sure what will influence the ratings of individual homemade student rating forms that have not been designed in accordance with professional psychometric standards.

Isn't it generally easier to get good ratings in higher level courses? The answer to this question is "yes." Actually, freshmen tend to rate a course more harshly than sophomores, sophomores more harshly than juniors, juniors more harshly than seniors, and seniors more harshly than graduate students. The majority of studies on this issue tend to support this belief. Aleamoni and Hexner (1980) cited 18 investigators who reported that graduate students and/or upper division students tended to rate instructors more favorably than did lower division students. They also cited eight investigators who reported no significant relationship between student status (freshman, sophomore, etc.) and ratings assigned to instructors.

Isn't it true that students who are required to take a course tend to rate the course more harshly than those taking it as an elective? Interestingly, the answer to this question is "yes." The bulk of the literature tends to support this conclusion. Several investigators have found that students who are required to take a course tend to rate it lower than students who elect to take it (Cohen & Humphreys, 1960; Gillmore & Brandenburg, 1974; Pohlmann, 1975). This finding is also supported by Gage (1961) and Lovell and Haner (1955) who found that instructors of elective courses were rated significantly higher than instructors of courses considered required by the students. In contrast, Heilman and Armentrout (1936) and Hildebrand, Wilson, and Dienst (1971) reported no differences between students' ratings of required courses and elective courses.

Isn't there a gender bias in student ratings? Don't female faculty tend to get lower ratings than male faculty? The literature on this question is not as plentiful as it is on other issues in student rating research, but the research that has been done appears to be fairly consistent in answering the question as "no." Earlier research provided conflicting results relating the gender of the student to student evaluations of instruction. Aleamoni and Thomas (1980), Doyle and Whitely (1974), Goodhartz (1948), and Isaacson, McKeachie, Milholland, Lin, Hofeller, Baerwaldt, et al. (1964) reported no differences between faculty ratings made by male and female students. In addition, Costin et al. (1971) cited seven studies that reported no differences in overall ratings of instructors made by male and female students or in the ratings received by male and female instructors. Conversely, Bendig (1952) found female students to be more critical of male instructors than their male counterparts. Aleamoni and Hexner (1980) cited five studies which reported female students rated instructors higher on some subscales of instructor evaluation forms than did male students. (See Basow & Howe, 1987; Basow & Silberg, 1987; Feldman, 1992, 1993; Ferber & Huber, 1975; Goodwin & Stevens, 1993; Kaschak, 1978.) Walker (1969) found that female students rated female instructors significantly higher than they rated male instructors. This finding was confirmed by Centra and Gaubatz (2000), but they concluded that even though the differences in ratings given by male and female students were statistically significant, they should not make much difference in personnel decisions. Thus, although it appears that female students may tend to rate female instructors higher than they rate male instructors, overall the literature does not indicate a consistent difference in the student ratings received by male and female instructors.

Isn't it more difficult for math and science faculty to get good ratings? Surprisingly the answer to this question is "yes." This does not necessarily mean that all math and science faculty are poor teachers. Rather, the normative data gathered on thousands of courses show that courses in the fields of math and science tend to get lower ratings than those in the humanities. No definitive reason for this tendency has yet been determined, but we do know that different disciplines produce different rating norms. That is, student ratings can be affected by the discipline being taught. It is important, therefore, in interpreting student rating results that we do so by using the appropriate norm group.

Isn't it true that the only faculty who are really qualified to teach or evaluate their peers' teaching are those who are actively involved in conducting research in their field? The answer to this question is "no." At one time it was believed that good instruction and good research were so closely allied that it was unnecessary to evaluate them independently (Borgatta, 1970; Deming, 1972). Some studies have found weak positive correlations between research productivity and teaching effectiveness (Maslow & Zimmerman, 1956; McDaniel & Feldhusen, 1970; McGrath, 1962; Riley, Ryan, & Lipschitz, 1950; Stallings & Singhal, 1968). On the other hand, Aleamoni and Yimer (1973), Guthrie (1949, 1954), Hayes (1971), Linsky and Straus (1975), and Voeks (1962) found no significant relationship between instructors' research productivity and students' ratings of their teaching effectiveness. One study

(Aleamoni & Yimer, 1973) also reported no significant relationship between instructors' research productivity and colleagues' ratings of their teaching effectiveness. More recently Hattie and Marsh (1996) and Marsh and Hattie (2002) have shown rather conclusively that the relationship between research productivity and teaching effectiveness is effectively zero. It is ironic that those individuals who apparently value research highly but persist in believing that only those involved in research can be effective teachers apparently do not read the literature on this issue.

Don't students have to be away from the course, and possibly the college, for several years before they are able to make accurate judgments about the instructor and instruction? The answer to this question is a qualified "no." This very popular belief is continuously bolstered by anecdotes passed from teacher to teacher. However, conducting research on this issue has proven to pose certain problems. For example, it is very difficult to obtain a comparative and representative sample in longitudinal follow-up studies. The sampling problem is further compounded by the fact that almost all student attitudinal data relating to a course or instructor are gathered anonymously, thus no true pre- and post-test research designs are possible. Most studies in this area, therefore, have relied on surveys of alumni and/or graduating seniors. Early studies by Drucker and Remmers (1951) showed that alumni who have been out of school 5 to 10 years rated instructors much the same as students currently enrolled. More recent evidence by Aleamoni and Yimer (1974), Marsh (1977), Marsh and Overall (1979), and McKeachie, Lin, and Mendelson (1978) further substantiated the earlier findings. Thus, the literature up to this point leads to the conclusion that this popularly held belief is not generally true.

Isn't it true that the size of the class affects student ratings? The answer to this question appears to be "no." Taken in its entirety, the literature has not indicated consistent relationship between class size and student ratings. Aleamoni and Hexner (1980) cited eight studies that indicate that class size has an effect on student ratings, and seven others that found no relationship between class size and ratings. Some investigations have also reported curvilinear relationships between class size and student ratings (Gage, 1961; Kohlan, 1973; Lovell & Haner, 1955; Marsh, Overall, & Kesler, 1979; Pohlmann, 1975; Wood, Linsky, & Straus, 1974). There does appear to be a substantial body of literature that student ratings are positively correlated with student learning, as measured by student performance on standardized final exams (Cohen, 1981, 1986; d'Apollonia & Abrami, 1988; McCallum, 1984). Also, there is evidence that students tend to rate most highly those courses in which they learn the most (Centra, 1977) with students who perceive the instructor to be well prepared and highly organized achieving the highest gains in learning (Pascarella, 2001). To the degree that faculty teaching courses of smaller size appear to the student to be better organized and prepared, or to the degree that students learn more in smaller classes, there may be some relationship between class size and student ratings. However, taken in its entirety, the literature does not support the position that a consistent relationship between class size (alone) and student ratings of any sort exists.

It is interesting to note, however, that most large, required courses tend to be offered early in the curriculum of a college. Thus, most of the large, required courses may be offered in the freshman and sophomore years, just the time when students tend to rate their teachers most harshly. It would be easy to conclude from personal experience with such courses that the problem lies with the size of the course when, in fact, the research indicates it is the level (freshman, sophomore, etc.) and the fact that the course is required that are the factors which contribute to generally lower student ratings.

Does the time of the day the course is taught affect student ratings? The answer to this question is "we don't think so." There hasn't been much research on this question, but the limited amount of research in this area (Feldman, 1978; Guthrie, 1954; Yongkittikul, Gillmore, & Brandenburg, 1974) indicates that the time of day the course is offered does not influence student ratings.

Do majors in a course rate it differently than nonmajors? The answer to this question is "no." Although there is only a limited amount of research in this area (Aleamoni & Thomas, 1980; Cohen & Humphreys, 1960; Null & Nicholson, 1972; Rayder, 1968), all studies indicate that there are no significant differences and no significant relationships between student ratings and whether students were majors or nonmajors.

Does the rank of the instructor (instructor, assistant professor, associate professor, or professor) affect student ratings? The answer to this question is "not really." The literature on this issue, in general, does not support the idea that faculty of higher professorial rank get higher student ratings, because no consistent relationship between faculty rank and student ratings has been found. Although some investigators reported that instructors of higher rank receive higher student ratings (Clark & Keller, 1954; Downie, 1952, Gage 1961; Guthrie, 1954; Walker, 1969), others reported no significant relationship between instructor rank and student ratings (Aleamoni & Graham, 1974; Aleamoni & Thomas, 1980; Aleamoni & Yimer, 1973; Linsky & Straus, 1975; Singhal, 1968).

Conflicting results have also been found when comparing teaching experience to student ratings. Rayder (1968) reported a negative relationship, whereas Heilman and Armentrout (1936) found no significant relationship.

Why bother with 20 or 30 questions on a student rating form? Can't we just use single general items as accurate measures of instructional effectiveness? The use of single general items on student rating forms has been popular for some time. These items can be very reliable but do not provide information that is as useful as that produced by the multiple-item student rating form. The limited amount of research in this area (Aleamoni & Thomas, 1980; Burdsal & Bardo, 1986) indicates that there is a low relationship between single general items and specific items and that the single general items had a much higher relationship to descriptive variables (gender, status, required-elective, etc.) than did the specific items. These findings suggest that the use of single general items should be avoided, especially for tenure, promotion, or salary considerations. In any case, single general items provide little useful diagnostic information for use in faculty development efforts compared to multiple items that measure specific components of teaching performance.

What good are student ratings in efforts to improve instruction? The answer to this question is "under the right conditions, student ratings can be very helpful in assisting faculty to improve their instruction." Studies by Braunstein, Klein, and Pachla (1973), Centra (1973a), and Miller (1971) were inconclusive with respect to the effect of feedback at midterm to instructors whose instruction was again evaluated at the end of the term. However, Marsh, Fleiner, and Thomas (1975), Overall and Marsh (1979), and Sherman (1978) reported more favorable ratings from and improved learning by students by the end of the term. In order to determine if a combination of a printed report of the results and personal consultations would be superior to providing only a printed report of results, Aleamoni (1978b), McKeachie (1979), and Stevens and Aleamoni (1985) found that instructors significantly improved their ratings when personal consultations were provided. The key finding that emerges here is that student ratings can be used to improve instruction if used as part of a personal consultation between the faculty member and a faculty development resource person.

Summary

Given the use of a well-designed, valid, and reliable student rating form, the literature indicates that:

- Faculty cannot buy good ratings by giving easy grades.
- Teaching a small class does not automatically guarantee high student ratings, nor does teaching a large class automatically guarantee low ratings.
- Lower level students (freshman, sophomore) do tend to rate more harshly than upper level students (seniors, graduate students).
- Students in required courses tend to rate their instructors more harshly than students in elective courses.
- Math and science courses tend to be rated more harshly than courses in the humanities.
- There is no overall gender bias in student ratings (i.e., one gender of faculty does not consistently get higher, or lower, ratings than the other).
- The time of day a class is taught does not affect the student ratings of the instructor (i.e., students do not automatically rate faculty lower in classes taught early in the morning or right after lunch).
- Student ratings can be quite helpful in instructional improvement efforts when used as part of a faculty development program that includes personal consultations.

Obviously the proper interpretation of student ratings must take a variety of issues into account. The important conclusion to be drawn here is the necessity to systematically incorporate research findings into the interpretation of student ratings in a comprehensive faculty evaluation system. See Chapter 14, especially the section on the Course Instructor Evaluation Questionnaire (CIEQ), for ways professionally developed, commercially available student rating form services address these issues.

References

Aleamoni, L. M. (1976). Typical faculty concerns about student evaluation of instruction. *National Association of Colleges and Teachers of Agriculture Journal, 20*(1), 16–21.

Aleamoni, L. M. (1978a). Development and factorial validation of the Arizona Course/Instructor Evaluation Questionnaire. *Educational and Psychological Measurement, 38*(6), 1063–1067.

Aleamoni, L. M. (1978b). The usefulness of student evaluations in improving college teaching. *Instructional Science, 7*(1), 95–105.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*(2), 153–166.

Aleamoni, L. M., & Graham, M. H. (1974). The relationship between CEQ ratings and instructor's rank, class size, and course level. *Journal of Educational Measurement, 11*(3), 189–202.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science, 9*(1), 67–84.

Aleamoni, L. M., & Spencer, R. E. (1973). The Illinois Course Evaluation Questionnaire: A description of its development and a report of some of its results. *Educational and Psychological Measurement, 33*, 669–684.

Aleamoni, L. M., & Thomas, G. S. (1980). Differential relationships of student, instructor, and course characteristics to general and specific items on a course questionnaire. *Teaching of Psychology, 7*(4), 233–235.

Aleamoni, L. M., & Yimer, M. (1973). An investigation of the relationship between colleague rating, student rating, research productivity, and academic rank in rating instructional effectiveness. *Journal of Educational Psychology, 64*(3), 274–277.

Aleamoni, L. M., & Yimer, M. (1974). *Graduating Senior Ratings Relationship to Colleague Rating, Student Rating, Research Productivity and Academic Rank in Rating Instructional Effectiveness* (Research Report No. 352). Urbana, IL: University of Illinois, Office of Instructional Resources, Measurement and Research Division.

Arreola, R. A. (1983). Students can distinguish between personality and content/organization in rating teachers. *Phi Delta Kappan, 65*(3), 222–223.

Basow, S. A., & Howe, K. G. (1987). Evaluations of college professors: Effects of professors' sex-type, and sex, and students' sex. *Psychological Reports, 60*, 671–678.

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*(3), 308–314.

Beatty, M. J., & Zahn, C. J. (1990). Are student rating of communication instructors due to "easy" grading practices? An analysis of teacher credibility and student-reported performance levels. *Communication Education, 39*(4), 275–282.

Bendig, A. W. (1952). A preliminary study of the effect of academic level, sex, and course variables on student rating of psychology instructors. *Journal of Psychology, 34*, 2–126.

Benz, Z., & Blatt, S. J. (1995). Factors underlying effective college-teaching: What students tell us. *Mid-Western Educational Researcher, 8*(1), 48–54.

Borgatta, E. F. (1970). Student ratings of faculty. *American Association of University Professors Bulletin, 56*, 6–7.

Braunstein, D. N., Klein, G. A., & Pachla, M. (1973). Feedback, expectancy, and shifts in student ratings of college faculty. *Journal of Applied Psychology, 58*(2), 254–258.

Burdsal. C. A., & Bardo, J. W. (1986). Measuring students' perceptions of teaching: Dimensions of evaluation. *Educational and Psychological Measurement, 56*, 63–79.

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. In. P. Seldin & Associates, *Changing practices in evaluating teaching* (pp. 25–44). Bolton, MA: Anker.

Centra, J. A. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal, 14*(1), 17–24.

Centra, J. A. (1973a). Effectiveness of student feedback in modifying college instruction. *Journal of Educational Psychology, 65*(3), 395–401.

Centra, J. A. (1973b). The student as godfather? The impact of student ratings on academia. In A. L. Sockloff (Ed.), *Proceedings of the First Invitational Conference on Faculty Effectiveness as Evaluated by Students.* Philadelphia, PA: Temple University Measurement and Research Center.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 70*(1), 17–23.

Clark, K. E., & Keller, R. J. (1954). Student ratings of college teaching. In R. A. Eckert & R. J. Keller (Eds.), *A university looks at its program: The report of the University of Minnesota Bureau of Institutional Research, 1942–1952.* Minneapolis, MN: University of Minnesota Press.

Cohen, P.A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multi-section validity studies. *Review of Educational Research, 51*(3), 281–309.

Cohen, P. A. (1986, April). *An updated and expanded meta-analysis of multisection validity studies.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Cohen, J., & Humphreys, L.G. (1960). *Memorandum to faculty.* Unpublished manuscript, University of Illinois Department of Psychology.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research, 41*(5), 511–535.

d'Apollonia, S., & Abrami, P.C. (1988, April). The literature on student ratings of instruction: Yet another meta-analysis. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Deming, W. E. (1972). Memorandum on teaching. *The American Statistician, 26*(1), 47.

Downie, N. W. (1952). Student evaluation of faculty. *Journal of Higher Education, 23,* 495–496, 503.

Doyle, K. O., & Whitely, S. E. (1974). Student ratings as criteria for effective teaching. *American Educational Research Journal, 11*(3), 259–274.

Drucker, A. J., & Remmers, H. H. (1951). Do alumni and students differ in their attitudes toward instructors? *Journal of Educational Psychology, 42*(3), 129–143.

Dukes, R. L., & Victoria, G. (1989). The effects of gender, status, and effective teaching on the evaluation of college instruction. *Teaching Sociology, 17*(4), 447–457.

Everly, J. C., & Aleamoni, L. M. (1972). The rise and fall of the advisor ... students attempt to evaluate their instructors. *Journal of the National Association of Colleges and Teachers of Agriculture, 16*(2), 43–45.

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

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*(6), 583–645.

Feldman, K. A. (1992). College students' views of male and female college teachers: Part I—Evidence from the social laboratory and experiments. *Research in Higher Education, 33*(3), 317–375.

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II—Evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34*(2), 151–211.

Ferber, M. A., & Huber, J. A. (1975). Sex of student and instructor: A study of student bias. American Journal of Sociology, 80, 949–963.

Frey, P. W. (1973). Student ratings of teaching: Validity of several rating factors. *Science, 182,* 83–85.

Frey, P. W. (1978). A two-dimensional analysis of student ratings of instruction. *Research in Higher Education, 9*(1), 69–91.

Gage, N. L. (1961). The appraisal of college teaching: An analysis of ends and means. *Journal of Higher Education, 32,* 17–22.

Gessner, P. K. (1973). Evaluation of instruction. *Science, 180,* 566–569.

Gillmore, G. M. (1973). *Estimates of reliability coefficients for items and subscales of the Illinois Course Evaluation Questionnaire* (Research Report No. 341). Urbana, IL: University of Illinois Office of Instructional Resources, Measurement, and Research Division.

Gillmore, G. M., & Brandenburg, D. C. (1974). *Would the proportion of students taking a class as a requirement affect the student rating of the course?* (Research Report No. 347). Urbana, IL: University of Illinois Office of Instructional Resources, Measurement, and Research Division.

Goodhartz, A. S. (1948). Student attitudes and opinions relating to teaching at Brooklyn College. *School and Society, 68,* 345–349.

Goodwin, L. D., & Stevens, E. A. (1993). The influence of gender on university faculty members' perceptions of "good" teaching. *Journal of Higher Education, 64*(2), 166–185.

Grush, J. E., & Costin, F. (1975). The student as consumer of the teaching process. *American Educational Research Journal, 12,* 55–66.

Guthrie, E. R. (1949). The evaluation of teaching. *Educational Record, 30,* 109–115.

Guthrie, E. R. (1954). *The evaluation of teaching: A progress report.* Seattle, WA: University of Washington.

Hattie, J., & Marsh, H. W. (1996). The relationship between research and teaching—A meta-analysis. *Review of Educational Research, 66*(4), 507–542.

Hayes, J. R. (1971). Research, teaching, and faculty fate. *Science, 172,* 227–230.

Heilman, J. D., & Armentrout, W. D. (1936). The rating of college teachers on ten traits by their students. *Journal of Educational Psychology, 27,* 197–216.

Hildebrand, M., Wilson, R. C., & Dienst, E. R. (1971). *Evaluating university teaching.* Berkeley, CA: University of California Center for Research and Development in Higher Education.

Hogan, T. P. (1973). Similarity of student ratings across instructors, courses, and time. *Research in Higher Education, 1*(2), 149–154.

Isaacson, R. L., McKeachie, W. J., Milholland, J. E., Lin, Y. G., Hofeller, M., Baerwaldt, J. W., & Zinn, K. L. (1964). Dimensions of student evaluations of teaching. *Journal of Educational Psychology, 55,* 344–351.

Johannessen, T. A., & Associates. (1997). What is important to students? Exploring dimensions in their evaluations of teachers. *Scandinavian Journal of Educational Research, 41*(2), 165–177.

Kaschak, E. (1978). Sex bias in student evaluations of college professors. *Psychology of Women Quarterly, 2*(3), 235–243.

Kohlan, R. G. (1973). A comparison of faculty evaluations early and late in the course. *Journal of Higher Education, 44,* 587–597.

Krehbiel, T. C., & Associates. (1997). Using student disconfirmation as a measure of classroom effectiveness, Journal of Educational Psychology, 82(2), 224–229.

Linsky, A. S., & Straus, M. A. (1975). Student evaluations, research productivity, and eminence of college faculty. *Journal of Higher Education, 46*(1), 89–102.

Lovell, G. D., & Haner, C. F. (1955). Forced-choice applied to college faculty rating. *Educational and Psychological Measurement, 15,* 291–304.

Macdonald, A. (1987). Student views on excellent courses. *Agricultural Education Magazine, 60*(3), 19–22.

Marlin, J. W., Jr. (1987). Student perception of end-of-course evaluations. *Journal of Higher Education, 58*(6), 704–716.

Marsh, H. W. (1977). The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. *American Educational Research Journal, 14*(4), 441–447.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*(5), 707–754.

Marsh, H. W., & Bailey, M. (1993). Multidimensionality of students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education, 64*(1), 1–18.

Marsh, H. W., Fleiner, H., & Thomas, C. S. (1975). Validity and usefulness of student evaluations of instructional quality. *Journal of Educational Psychology, 67*(6), 883–889.

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*(2), 139–147.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Class size, students' evaluations, and instructional effectiveness. *American Educational Research Journal, 16*(1), 57–69.

Marsh, H. W., & Hattie, J. (2002). The relationship between research productivity and teaching effectiveness. *The Journal of Higher Education, 73*(5), 603–641.

Maslow, A. H., & Zimmerman, W. (1956). College teaching ability, scholarly activity, and personality. *Journal of Educational Psychology, 47,* 185–189.

McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in tenure decisions. *Research in Higher Education, 21*(2), 150–158

McDaniel, E. D., & Feldhusen, J. F. (1970). Relationships between faculty ratings and indexes of service and scholarship. *Proceedings of the 78th Annual Convention of the American Psychological Association, 5,* 619–620.

McGrath, E. J. (1962). Characteristics of outstanding college teachers. *Journal of Higher Education, 33,* 148–152.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe, 65*(6), 384–397.

McKeachie, W. J., Lin, Y. G., & Mendelson, C. N. (1978). A small study assessing teacher effectiveness: Does learning last? *Contemporary Educational Psychology, 3*(4), 352–357.

Menges, R. J. (1973). The new reporters: Students rate instruction. In C. R. Pace (Ed.), *New directions in higher education: No. 4. Evaluating learning and teaching* (pp. 59–75). San Francisco, CA: Jossey-Bass.

Merriam-Webster OnLine. (2005–2006). *Valid.* Retrieved June 28, 2006, from http://www.m-w.com/dictionary/valid

Miller, M. T. (1971). Instructor attitudes toward, and their use of, student ratings of teachers. *Journal of Educational Psychology, 62*(3), 235–239.

Null, E. J., & Nicholson, E. W. (1972). Personal variables of students and their perception of university instructors. *College Student Journal, 6,* 6–9.

Overall, J. U., & Marsh, H. W. (1979). Midterm feedback from students. Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of Educational Psychology, 71,* 856–865.

Pascarella, E. T. (2001). Cognitive growth in college: Surprising and reassuring findings from The National Study of Student Learning. *Change, 33*(6), 21–27.

Perry, R. P., Abrami, P. C., & Leventhal, L. (1979). Educational seduction: The effect of instructor expressiveness and lecture content on student ratings and achievement. *Journal of Educational Psychology, 71*(1), 107–116.

Pohlmann, J. T. (1975). A multivariate analysis of selected class characteristics and student ratings of instruction. *Multivariate Behavioral Research, 10*(1), 81–91.

Rayder, N. F. (1968). College student ratings of instructors. *Journal of Experimental Education, 37,* 76–81.

Riley, J. W., Ryan, B. F., & Lipschitz, M. (1950). *The student looks at his teacher.* New Brunswick, NJ: Rutgers University Press.

Rodabaugh, R. C., & Kravitz, D. A. (1994). Effects of procedural fairness on student judgments of professors. *Journal on Excellence in College Teaching, 5*(2).

Seldin, P. (1993). How colleges evaluate professors: 1983 vs. 1993. *AAHE Bulletin, 46*(2), 6–8, 12.

Shepherd, G. J. & Trank, D. M. (1989). Individual difference in consistency of evaluation: Student perceptions of teacher effectiveness. *Journal of Research and Development in Education, 22*(3), 45–52.

Sherman, T. M. (1978). The effects of student formative evaluation of instruction on teacher behavior. *Journal of Educational Technology Systems, 6,* 209–217.

Singhal, S. (1968). *Illinois course evaluation questionnaire items by rank of instructor, sex of the instructor, and sex of the student* (Research Report No. 282). Urbana, IL: University of Illinois Office of Instructional Resources, Measurement, and Research Division.

Stallings, W. M., & Singhal, S. (1968). *Some observations on the relationships between productivity and student evaluations of courses and teaching* (Research Report No. 274). Urbana, IL: University of Illinois Office of Instructional Resources, Measurement, and Research Division.

Stevens, J. J., & Aleamoni, L. M. (1985). The use of evaluative feedback for instructional improvement: A longitudinal perspective. *Instructional Science, 13,* 285–304.

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

Thorndike, R.L. & Hagen, E. (1969). *Measurement and evaluation in psychology and education* (3rd ed.). New York, NY: John Wiley & Sons.

Tollefson, N., Chen, J. S., & Kleinsasser, A. (1989). The relationship of students' attitudes about effective teaching to students' ratings of effective teaching. Educational and Psychological Measurement, 49(3), 529–536.

Voeks, V. W. (1962). Publications and teaching effectiveness. *Journal of Higher Education, 33,* 212–218.

Walker, B. D. (1969). An investigation of selected variables relative to the manner in which a population of junior college students evaluate their teachers. *Dissertation Abstracts, 29*(9–B), 3474.

Ware, J. E., Jr., & Williams, R. G. (1977). Discriminant analysis of student ratings as a means of identifying lecturers who differ in enthusiasm or information giving. *Educational and Psychological Measurement, 37*(3), 627–639.

Waters, M., Kemp, E., & Pucci, A. (1988). High and low faculty evaluations: Descriptions by students. *Teaching of Psychology, 15,* 203–204.

Wood, K., Linsky, A. S., & Straus, M. A. (1974). Class size and student evaluation of faculty. *Journal of Higher Education, 45*(7), 524–534.

Yongkittikul, C., Gillmore, G. M., & Brandenburg, D. C. (1974). *Does the time of course meeting affect course ratings by students?* (Research Report No. 346). Urbana, IL: University of Illinois Office of Instructional Resources, Measurement, and Research Division.

Portions of this chapter appeared in the following article. Used by permission.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*(2), 153–166.