IDEA PAPER NO. 32

center for FACULTY EVALUATION & DEVELOPMENT
Division of Continuing Education
Kansas State University

September, 1995

Student Ratings of Teaching: The Research Revisited

William E. Cashin
Kansas State University

> Negative attitudes toward student ratings are especially resistant to change, and it seems that faculty and administrators support their belief in student-rating myths with personal and anecdotal evidence, which [for them] outweighs empirically based research evidence.
> (Cohen, 1990, p. 124–125)

There are now more than 1,500 references dealing with research on student evaluations of teaching. IDEA Paper No. 20, *Student Ratings of Teaching: A Summary of the Research* (Cashin, 1988) attempted to briefly summarize the research from 1971 to 1988. This paper is an update of that paper and repeats much of its content. No major study published since then has substantively changed that paper's conclusions, but several studies or reviews of the literature provide modifications or further support for its conclusions.

This paper will attempt to summarize the conclusions of the major reviews of the student rating literature from Costin, Greenough, and Menges (1971) to the present. That literature is extensive and complex. Obviously, a paper this brief can offer only broad, general conclusions and very limited citations. Interested readers are encouraged to consult the various reviews and their individual references for details. For readers with less time, both Braskamp and Ory (1994) and Centra (1993) have chapters summarizing the student rating research; see also Davis (1993) and McKeachie (1994).

The ERIC descriptor for student ratings is "student evaluation of teacher performance". I suggest that the term "student ratings" is preferable to "student evaluations." "Evaluation" has a definitive and terminal connotation; it suggests that we have an answer. "Rating" implies that we have data which need to be interpreted. Using the term "rating" rather than "evaluation" helps to distinguish between the people who provide the information (sources of data) and the people who interpret it in combination with other sources of data (evaluators).

Viewing student ratings as data rather than as evaluations may also help to put them in proper perspective. Writers on faculty evaluation are almost universal in recommending the use of *multiple* sources of data. No single source of data—including student rating data—provides sufficient information to make a valid judgment about overall teaching effectiveness. Further, there are important aspects of teaching that students are *not competent* to rate (see IDEA Paper No. 21, *Defining and Evaluating College Teaching*, Cashin, 1989, for details.)

Multidimensionality

There have been a number of factor analytic studies (see Abrami & d'Apollonia, 1990; Feldman, 1976b; Kulik & McKeachie, 1975; and Marsh & Dunkin, 1992, for details) that conclude that student rating forms are multidimensional, i.e., that they measure *several different aspects* of teaching. Put another way, ***no* single student rating item, nor set of *related* items, will be useful for *all* purposes.**

Both Centra (1993) and Braskamp and Ory (1994) identify six factors commonly found in student rating forms:

1. Course organization and planning
2. Clarity, communication skills
3. Teacher student interaction, rapport
4. Course difficulty, workload
5. Grading and examinations
6. Student self-rated learning

Marsh's (1984) SEEQ (Students' Evaluations of Educational Quality) form has nine dimensions: learning/value, enthusiasm, organization, group interaction, individual rapport, breadth of coverage, exams/grades, assignments, and workload. Other student rating forms have items measuring some or all of the above dimensions. In several of his reviews of the literature, Feldman (1976b, 1983, 1984, 1987, and 1988) categorized student ratings items—and gave examples—into as many as 22 different logical dimensions. In a more recent review, Feldman (1989b) identified 28 dimensions. When interpreting student rating data, **we must distinguish among the various items and their dimensions to insure that all of the appropriate**

dimensions are rated. Averaging *dissimilar* items is *not* appropriate.

Although there is general agreement that student ratings are multidimensional, and that various dimensions should be used *when their purpose is to improve teaching*, there is disagreement about how many, or which, dimensions should be used *for personnel decisions*. In several articles Abrami (e.g., 1989a; and Abrami & d'Apollonia, 1991) suggested that **one or a few global or summary type items might provide sufficient student rating data for personnel decisions**. Centra (1993) and Braskamp and Ory (1994) make a similar recommendation. Cashin and Downey (1992) tested this using the IDEA Overall Evaluation measure as the criterion of teaching effectiveness. Each of three global items—individually—accounted for at least 50% of the variance in the criterion measure: overall instructor effectiveness, 54%; overall course worth, 60%; overall amount learned, 69%. However—contrary to their hypothesis—controlling for the students' motivation to take the course, the size of the class, or the difficulty of the subject matter, did *not* add significantly to the amount of variance explained. Marsh (1994) had some reservations about the way the IDEA Overall Evaluation measure was calculated and he generated four variations that he considered improvements. However, Cashin, Downey, and Sixbury (1994)—using each of Marsh's four variations as the criterion measure—obtained the same results as the original study: each of the global items accounted for at least 50% of the variance in each of Marsh's criterion measures, and the control items added little.

Reliability

In the educational measurement literature, reliability covers consistency, stability, and generalizability of items. For student rating items, reliability refers most often to **consistency** or interrater agreement (i.e., within a given class do the students tend to give similar ratings on a given item). **Reliability varies depending upon the number of raters, i.e., the more raters, the more reliable**. For example, with the IDEA system (Sixbury & Cashin, 1995a), the median reliabilities (intraclass correlations) for the 38 items are:

for 10 raters, .69
for 15 raters, .83
for 20 raters, .83
for 30 raters, .88
for 40 raters, .91

Similar or higher reliabilities are typically found with other well-designed forms, i.e., forms developed with the assistance of someone knowledgeable about educational measurement. **As a rule of thumb, I recommend that items with fewer than ten raters (reliabilities below .70), be interpreted with *particular caution*.**

Stability is concerned with agreement between raters *over time*. In general, ratings of the same instructor tend to be similar over time (Braskamp & Ory, 1994; Centra, 1993). For example, a longitudinal study (Overall & Marsh, 1980) compared end-of-course ratings with ratings by the same students years later (at least one year after graduation). The average correlation was .83.

Generalizability is concerned with how confident we can be that our data accurately reflect the instructor's *general* teaching effectiveness, not just how effective he or she was in that particular course that term. A study conducted by Marsh (1982) illustrates the question. He studied data from 1,364 courses, dividing them into four categories: the same instructor teaching the same course but in different terms, the same instructor teaching a different course, different instructors teaching the same course, and different instructors teaching different courses. This permitted him to study the differential effects of the instructor and of the course. He then correlated student ratings in the four different categories, separating items related to the instructor (e.g., enthusiasm, organization, discussion) from background items (e.g., student's reason for taking the course, workload). The average correlations are shown below; the correlations in parentheses are for the background items.

	Same Course	Different Course
Same Instructor	.71 (.69)	.52 (.34)
Different Instructor	.14 (.49)	.06 (.21)

The *instructor-related* correlations were higher for the same instructor, even when teaching a different course. The correlations for the *background* items (in parentheses)—more tied to the course than the instructor—were higher for the same course. Marsh concluded that **the instructor, not the course, is the primary determinant of the student rating items**. Marsh's results are comparable to other generalizability studies (Gillmore, Kane, & Naccarato, 1978; and Hogan, 1973).

When making personnel decisions, we want to use the data to make judgments about the instructor's *general* teaching effectiveness. When considering student ratings (remembering that we need other kinds of information beyond student ratings), the following seem to be reasonable rules of thumb. If the instructor teaches only one course (e.g., part-time instructors), *consistent* ratings from two different terms may be sufficient. For most instructors, however, **use ratings from a *variety of courses*, for two or more courses from every term for at least two years, totaling at least five courses. If there are fewer than fifteen raters in any of the classes, data from additional classes are recommended.**

Validity

In educational measurement, the basic question concerning validity is: does the test measure what it is supposed to measure? For student ratings this translates into: **to what extent do student rating items measure some aspect of teaching effectiveness?** Unfortunately there is no agreed upon definition of

"effective teaching" nor any single, all-embracing criterion. The best that one can do is to try various approaches, collecting data that either support or contest the conclusion that student ratings reflect effective teaching.

Approach One—Student Learning

Theoretically, the best criterion of effective teaching is student learning. Other things being equal, the students of more effective teachers should learn more. A number of studies have attempted to study this hypothesis by comparing *multiple-section* courses. In the typical study, different instructors teach different sections of the same course, using the same syllabus and textbook, and most importantly using the same *external* final exam, i.e., an exam developed by someone *other* than the instructors. Cohen (1981) and Feldman (1989b) reviewed these studies. Using the students' grades on the external exam as the measure of student learning, they examined correlations between the exam grade and various student rating items. The average correlations are given below (1981–Cohen; 1989–Feldman):

Student ratings of	1981	1989
achievement or learning	.47	.46
overall course	.47	–
overall instructor	.44	–
teacher skill dimension	.50	–
–course preparation	–	.57
–clarity of objectives	–	.35
teacher structure dimension	.47	–
–understandableness	–	.56
–knowledge of subject	–	.34
teacher rapport dimension	.31	–
–availability	–	.36
–respect for students	–	.23
teacher interaction dimension	.22	–
–encouraging discussion	–	.36

Note on Interpreting Validity Correlations: *Earlier I suggested as a rule of thumb that* ***reliability*** *correlations of* ***at least*** *.70 (at least 10 raters) were desirable. However, in the social sciences* ***validity*** *correlations* ***above*** *.70 are unusual, especially if studying complex phenomena, such as student learning. As a rule of thumb, I suggest that student rating validity correlations between .00 and .29, even when statistically significant, are not practically useful. Correlations between .30 and .49 are practically useful. Correlations between .50 and .70 are very useful but are not common when studying complex phenomena.*

Using the above rule of thumb, the average correlations reported by Cohen (1981) and Feldman (1989b) are generally useful. These relationships tend to support the validity of student ratings because **the classes in which the students gave the instructor higher ratings tended to be the classes where the students learned more**, i.e., scored higher on the external exam. On the other hand, the correlations are far from perfect, in part because many of the variables that relate to students' learning will be related to *student* characteristics (e.g., motivation or ability), not to instructor characteristics.

Approach Two—Instructor's Self Ratings

Researchers have sought for a criterion of effective teaching that would be acceptable to faculty. One possibility is the self ratings of the instructor. In a review of the literature, Feldman (1989a) cites 19 studies which correlated instructor's self ratings with student ratings. The average correlation was .29. However, in one study (Marsh, Overall, & Kesler, 1979) instructors were asked to rate *two different* courses in order to see if the course the instructor rated higher was also rated higher by the students. The median correlation—*based on six factor scores* between the instructor's self ratings and the students' ratings—was .49. In a later report (Marsh & Dunkin, 1992) using nine factor scores, the median was .45. Such studies provide further support for the validity of the students' ratings.

Approach Three—The Ratings of Others

If one is willing to grant that the ratings of administrators, colleagues, alumni, and others have some validity—and, excepting alumni, that these ratings are *independent* of feedback from students—then student ratings share that validity.

Administrator's Ratings—Student ratings correlate with administrator's ratings, ranging from .47 to .62 (Kulik & McKeachie, 1975), but Feldman (1989a), using global items, found a lower average correlation of .39.

Colleague's Ratings—Student ratings correlate with colleague's ratings, .48 to .69 (Kulik & McKeachie, 1975); Feldman (1989a) found an average of .55. Marsh and Dunkin (1992) question the usefulness of colleague's ratings *based on classroom visitation* because such ratings tend to be unreliable.

Some faculty question whether the students have an appropriate conception of what effective teaching is. In a review of 31 studies, Feldman (1988) found that the students' view of effective teaching was very similar to the faculty's view (average correlation equalled .71). There were some differences in emphasis between the two groups. Students tended to place more weight on the instructor being interesting, having good speaking skills, and being available to help; students also focused more on the outcomes of instruction, e.g., what they learned. Faculty placed relatively more weight on intellectual challenge, motivating students, setting high standards, and fostering student self-initiated learning.

Alumni Ratings—Student ratings correlate with alumni ratings, .40 to .75 (Overall & Marsh, 1980; Braskamp & Ory, 1994). Feldman (1989a) found an average correlation of .69. This belies the conventional wisdom that the students will come to appreciate our teaching after they get into the real world as working adults.

Trained Observers—A few studies have used external observers who were trained (see Feldman, 1989a, also Marsh & Dunkin, 1992). Reviewing five studies, Feldman found positive correlations with global student ratings (average correlation was .50). On a related issue, in another study (Murray, 1983) the median reliability for trained observers was .76. This suggests that peer ratings based on classroom observation would be reliable *if the observers were trained.*

Approach Four—Comparison with Student Comments

Some faculty question the value of student ratings but accept student written comments to open-ended questions. One study (Ory, Braskamp, & Pieper, 1980) of 14 classes found a correlation of .93 between a global instructor item and the students comments. A second study (Braskamp, Ory, & Pieper, 1981) of 60 classes found a correlation of .75. These studies suggest that, for personnel decisions, the information from student ratings overlaps considerably the information in student comments.

Approach Five—Possible Sources of Bias

One need not talk with faculty very long to be aware of their concern about possible biases in student ratings—about variables that correlate with student ratings. Some writers have suggested that bias be defined as anything *not under the control of the instructor.* Marsh (1984) argued *against* this definition because, for example, grading leniency—instructors giving higher grades than the students earned—would *not* be considered a bias using this definition. Marsh suggests that **bias in student ratings** should be restricted to **variables *NOT related to teaching effectiveness***. By this definition, the correlations between student ratings and class size, or the students' interest in the course are *not* biases because it is probable that students in small classes, or classes of students who are interested in the subject matter *actually do learn more.*

In IDEA Paper No. 20 (Cashin, 1988), I suggested an even narrower definition *when using ratings for personnel decisions or the instructor's improvement.* I suggested restricting bias to variables *not* a function of the *instructor's* teaching effectiveness. Thus, student motivation or class size might impact teaching effectiveness, but instructors should not be faulted if they were less effective teaching large classes of unmotivated students than their colleagues who were teaching small classes of motivated students. In this case, student motivation and class size, although related to teaching effectiveness, were *not* a function of the instructor's characteristics, but of student and course characteristics. Thus, they should be considered sources of bias, and should be controlled for by using appropriate comparative data. Feldman (1995, April) observed—accurately in my judgment—that such a definition of bias, while possibly acceptable, was not the usual definition and it served to confuse the literature. Marsh and Dunkin (1992)—considering that prior student interest in the subject matter is *not* a bias because it does impact teaching and learning—raise the question of "fairness" in comparing instructors teaching classes of interested students versus instructors teaching classes of uninterested students.

In the interest of clarity, rather than using "bias" in the restricted sense I did in the original paper, I will identify variables (when correlated with student ratings) that require control, especially when making personnel decisions.

Variables *Not* Requiring Control

Despite widespread faculty concern, the research has uncovered relatively few variables that correlate with student ratings *but are not* related to instructional effectiveness. Generally the following variables tend to show *little or no* relationship to student ratings:

A. *Instructor* variables *not* related to student ratings:

1) **age, and teaching experience**—in general age, and also years of teaching experience, are *not* correlated with student ratings. However, where small differences have been found, they tend to be negative, i.e., older faculty receive *lower* ratings (Feldman, 1983). Marsh and Hocevar (1991) point out that most of the studies have been cross-sectional, studying different cohorts of faculty to represent different age groups. In a longitudinal study they analyzed student ratings of the *same* instructors for as long as 13 years. They found *no* systematic changes over the years.

2) **gender of the instructor**—in a review of 14 *laboratory or experimental* studies, e.g., where students rated descriptions of fictitious teachers, Feldman (1992) found *no* differences in global ratings in the majority of studies, but in a few studies the male teachers received higher ratings. In a second review of 28 studies of *actual ratings of real* teachers reporting global ratings, he (Feldman, 1993) found a very slight average difference in favor of women teachers (r = .02). However, a few studies raised the question of whether women faculty had to do *more* of what was being rated (e.g., being available to students) to obtain the *same* ratings as men. In *a few other* studies there was a gender of student/gender of instructor interaction, i.e., female students rated female teachers higher, and male students rated male instructors higher.

3) **race**—Centra (1993) points out that there have been hardly any studies of the race of the instructor. He speculates that students of the same race as the instructor might rate the instructor higher. In a doctoral dissertation using IDEA, Li (1993) found no difference in the global ratings of Asian students compared to American students of their (presumably Caucasian) instructors.

4) **personality**—few personality traits tend to correlate with student ratings (Braskamp & Ory, 1994; Centra, 1993). In studies measuring personality using instructor's self report (e.g., personality inventories, self-description questionnaires), Feldman (1986) found only two (out of fourteen) traits that had average correlation with a global item that approached practical significant correlations. These traits were **positive self esteem** (r = .30), and **energy and enthusiasm** (r =

.27). *Note, I suggest that these two traits enhance the instructor's teaching effectiveness and so should* ***not*** *be controlled.* Murray, Rushton, and Paunonen (1990) found significantly different patterns of personality traits of psychology instructors teaching six different types of courses, e.g., introductory, graduate. They concluded that instructors tend to be differentially suited to different types of courses.

5) **research productivity**—has little correlation with student ratings (Centra, 1993). In his review of the literature, Feldman (1987) found the average correlation between research productivity and overall teaching effectiveness items to be .12. This very low correlation suggests that research productivity is indicative neither of good teaching nor bad teaching.

B. *Student* variables *not* related to student ratings:

1) **age of the student**—(Centra, 1993).
2) **gender of the student**—(Feldman, 1977, 1993), but sometimes there is a gender of student/gender of instructor interaction (see above under instructor variables).
3) **level of the student**—e.g., freshman (McKeachie, 1979).
4) **student's GPA**—(Feldman, 1976a).
5) **student's personality**—(Abrami, Perry, & Leventhal, 1982).

C. *Course* variables *not* related to student ratings:

1) **class size**—although there is a tendency for smaller classes to receive higher ratings, it is a very weak inverse association, i.e., smaller classes receive higher ratings, average r = -.09 (Feldman, 1984). The average correlation of class size for the 38 IDEA items is -.14 (Sixbury & Cashin, 1995a).
2) **time of day** when the course is taught—(Aleamoni, 1981; Feldman, 1978).

D. *Administrative* variables *not* related to student ratings:

1) **time during the term** when ratings are collected; any time during the second half seems to yield similar ratings—(Feldman, 1979).

Variables *Possibly* Requiring Control

The research cited above suggests that many variables suspected of biasing student ratings are *not* correlated with them to any practically significant degree. For the following variables, however, the research suggests that there are correlations—relationships—with student ratings that may require control.

A. *Instructor* variables *related* to student ratings:

1) **faculty rank**—regular faculty tend to receive higher ratings than graduate teaching assistants (Braskamp & Ory, 1994). This variable does *NOT* require control because regular faculty as a group tend to be more effective teachers than GTAs as a group.

2) **expressiveness**—the Dr. Fox effect (Naftulin, Ware, & Donnelly, 1973)—where a professional actor delivering little content received high ratings—suggests that student ratings may be more influenced by an instructor's style of presentation than by the substance of the content. The literature is complex (see Abrami, Leventhal, & Perry, 1982), but Marsh and Ware (1982) suggest that, especially in studies involving an incentive and a test, manipulations of instructor expressiveness primarily influences items related to instructor enthusiasm, and manipulation of content coverage primarily influences items related to instructor knowledge and student exam performance. Nevertheless, making the class interesting as well as informative helps students learn content. Expressiveness tends to enhance learning and does *NOT* require control.

B. *Student* variables *related* to student ratings:

1) **student motivation**—instructors are more likely to receive higher ratings in classes where students had a prior interest in the subject matter (Marsh & Dunkin, 1992), or were taking the course as an elective (Aleamoni, 1981; Braskamp & Ory, 1994; Centra, 1993; Feldman, 1978). The average correlation of the IDEA (Sixbury & Cashin, 1995a) motivation item, "I had a strong desire to take this course," with the other 37 items is .40. Marsh and Dunkin (1992) conclude that **reason for taking the course** (which overlaps with student motivation), also is related to student ratings. Higher ratings were received from students who took a course for general interest, or as a major elective; lower ratings were received when the course is being taken as a major requirement or a general education requirement. This variable *REQUIRES CONTROL.*

2) **expected grades**—there tend to be positive, but low correlations (.10 to .30) between students ratings and expected grades (Braskamp & Ory, 1994; Feldman, 1976a; Howard & Maxwell, 1980 and 1982; Marsh & Dunkin, 1992). Three possible hypotheses have been proposed for these correlations. One is the **validity hypothesis**—the students who learned more earn higher grades *and* give higher ratings (therefore, student ratings are valid). Another explanation is **grading leniency**—instructors giving higher grades than the students deserve receive higher ratings than they deserve. A third is based on **student characteristics**—some student characteristics, e.g., high motivation, lead to greater learning and, therefore, to higher grades *and* higher ratings. In two studies by Howard and Maxwell (1980 & 1982), which used IDEA data, they concluded that most of the correlation between expected grade and a global instructor item was accounted for by student (self-reported) learning—the validity hypothesis—and desire to take the course—a student characteristic. To control for the possibility of grade leniency, my recommendation is to have peers (faculty *knowledgeable in the subject matter*) review the course material, particularly exams, computer scored test results, graded samples of essays, projects, etc.; and judge whether grades are inflated.

C. *Course* variables *related* to student ratings:

1) **level of the course**—higher level courses, especially graduate courses, tend to receive higher ratings (Aleamoni, 1981; Braskamp & Ory, 1994; Feldman, 1978). However, the differences tend to be small. Regarding possible control, check to see if your

freshman/sophomore classes receive lower ratings than your junior/senior classes; similarly compare undergraduate with graduate classes. If yes, do the differences remain after controlling for student motivation and size? If yes, develop comparative data for the appropriate levels.

2) **academic field**—Feldman (1978) reviewed some studies showing that humanities and arts type courses receive higher ratings than social science type courses, which in turn receive higher ratings than math-science type courses. Others (Braskamp & Ory, 1994; Cashin, 1990; Centra, 1993; Marsh & Dunkin, 1992; and Sixbury & Cashin, 1995b) have found similar results. Although there is increasing evidence that ratings for different fields differ, it is not clear why. Cashin (1990) suggests six possible explanations. For example, if some fields are rated lower because they are more poorly taught, then these differences do *not* require control. On the other hand, if instructors in fields requiring more quantitative reasoning skills are rated lower because today's students are less competent in such skills—one of the hypotheses explaining why some fields are rated lower—then this should be controlled for.

3) **workload/difficulty**—these are correlated with student ratings (Centra, 1993; Marsh & Dunkin, 1992). However, contrary to faculty belief, they are correlated *positively*, i.e., students give *higher* ratings in difficult courses where they have to work hard. Although positive, the correlations are not large. For example, using the 38 IDEA items (Sixbury & Cashin, 1995a) the average correlations with the remaining 37 IDEA items are:

Amount of reading	.11
Amount of other (non reading) assignments	.16
Difficulty of subject matter	.15
Worked harder in this course	.29

These modest results support the validity of student ratings and the variables do *NOT* require control.

D. *Administrative* variables *related* to student ratings:

1) **non-anonymous ratings**—signed ratings tend to be higher (Braskamp & Ory, 1994; Centra, 1993; Feldman, 1979; Marsh & Dunkin, 1992). The hypothesis is that requiring students to sign their names inflates the ratings because some students are concerned about possible reprisals. Control: instruct the students *not* to sign their ratings.

2) **instructor present while students complete ratings**—these tend to be higher (Braskamp & Ory, 1994; Centra, 1993; Feldman, 1979; Marsh & Dunkin, 1992), possibly for the same reason as non-anonymous ratings. Control: have the instructor leave the room while the ratings are being completed and collected.

3) **purpose of the ratings**—some studies have found that if the directions say the ratings will be used for personnel decisions, the ratings tend to be higher than if they will be used only by the instructor for improvement (Braskamp & Ory, 1994; Centra, 1993; Feldman, 1979; Marsh & Dunkin, 1992). Speculation is that the students tend to be lenient if the data will be used by someone other than the instructor. Control: include in the standard directions the purpose(s) for which the ratings will be used. This will *not* eliminate the bias, but it will eliminate variations in ratings due to differences in student beliefs about their purpose.

Usefulness of Student Ratings

Many faculty will grant the usefulness of student ratings for personnel decisions, but question their usefulness for improvement, preferring to rely on students' open-ended comments. Cohen (1980) performed a meta-analysis of 17 studies of the effect of student-rating feedback on improving teaching. Receiving feedback about student ratings administered during the first half of the term was *positively* related to improving college teaching as measured by student ratings administered at the end of the term. Typically there were three groups. All groups had ratings administered during the first half of the semester and again at the end. That is all the first group received, i.e., no feedback. The second group received the student rating feedback, quantitative data, from the first student ratings. In addition to that, the third group received some kind of consultation (which varied across the different studies). Using the end-of-term ratings as the measure of improvement and setting the first group's mean ratings at the 50th percentile, Cohen presented the following data:

During term	End of Term
No student rating feedback =	50th %ile
Only student rating feedback =	58th %ile
Student rating feedback *plus consultation* =	74th %ile

Conclusion, if an institution really intends to use student ratings to improve teaching, it needs to provide some kind of consultation to the instructors.

Conclusion

There are probably more studies of student ratings than of all of the other data used to evaluate college teaching combined. Although one can find individual studies that support almost any conclusion, for a number of variables there are enough studies to discern trends. In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for evaluation. Nevertheless, student ratings are only one source of data about teaching and must be used in combination with multiple sources of data if one wishes to make a judgment about all of the components of college teaching. Further, student ratings are data that must be interpreted. We should not confuse a source of data with the evaluators who use student rating data—in combination with other kinds of data—to make their judgments about an instructor's teaching effectiveness.

References and Suggested Reading

Abrami, P. C. (1989a). How should we use student ratings to evaluate teaching? *Research in Higher Education, 30*, 221–227.

Abrami, P. C. (1989b). SEEQing the truth about student ratings of instruction. *Educational Researcher, 43*, 43–45.

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

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

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

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

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

Benton, S. E. (1982). *Rating college teachers: Criterion validity studies of student evaluation-of-instruction instruments.* AAHE-ERIC Higher Education Research Report, No. 1. Washington, DC: American Association for Higher Education.

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

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

Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research.* IDEA Paper No. 20. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development. (ERIC Document Reproduction Service No. ED 302 567) (Reprinted in K. A. Feldman & M. B. Paulsen (Eds.). (1994). *Teaching and learning in the college classroom* (pp. 531–541). Needham Heights, MA: Ginn Press).

Cashin, W. E. (1989). *Defining and evaluating college teaching.* IDEA Paper No. 21. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development. (ERIC Document Reproduction Service No. ED 339 731)

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

Cashin, W. E. (1992). Student ratings: The need for comparative data. *Instructional Evaluation and Faculty Development, 12*, 1–6.

Cashin, W. E., & Downey, R. G. (1992). Using global student ratings for summative evaluation. *Journal of Educational Psychology, 84*, 563–572.

Cashin, W. E., Downey, R. G., & Sixbury, G. R. (1994). Global and specific ratings of teaching effectiveness and their relation to course objectives: Reply to Marsh (1994). *Journal of Educational Psychology, 86*, 649–657.

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

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

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

Cohen, P. A. (1990). Bring research into practice. In M. Theall, & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice: New Directions for Teaching and Learning*, No. 43 (pp. 123–132). San Francisco: Jossey-Bass.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research, 8*, 511–535.

Davis, B. G. (1993). *Tools of teaching.* San Francisco: Jossey-Bass.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Feldman, K. A. (1995, April). *Some unresolved issues in studying instructional effectiveness and student ratings.* Invited address presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement, 15*, 1–13.

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

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

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

Kulik, J. A., & McKeachie, W. J. (1975). The evaluation of teachers in higher education. In F. N. Kerlinger (Ed.), *Review of research in education* (Vol. 3, pp. 210–240). Itasca, IL: F. E. Peacock.

Li, Y. (1993). *A comparative study of Asian and American students' perceptions of faculty teaching effectiveness at Ohio University.* Unpublished doctoral dissertation, Ohio University, Athens.

Marsh, H. W. (1982). The use of path analysis to estimate teacher and course effects in student rating's of instructional effectiveness. *Applied Psychological Measurement, 6*, 47–59.

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

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253–388.

Marsh, H. W. (1991a). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology, 83*, 285–296.

Marsh, H. W. (1991b). A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami and d'Apollonia (1991). *Journal of Educational Psychology, 83*, 416–421.

Marsh, H. W. (1994). Weighting for the right criteria in the IDEA System: Global and specific ratings of teaching effectiveness and their relation to course objectives. *Journal of Educational Psychology, 86*, 631–648.

Marsh, H. W., & Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.) *Higher education: Handbook of theory and research* (Vol. 8, pp. 143–233). New York: Agathon.

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

Marsh, H. W., & Overall, J. U. (1979). Long-term stability of students' evaluations: A note on Feldman's consistency and variability among college students in rating their teachers and courses. *Research in Higher Education, 10*, 139–147.

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

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

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

McKeachie, W. J. (1990). Research on college teaching: The historical background. *Journal of Educational Psychology, 82*, 189–200.

McKeachie, W. J. (1994). *Teaching tips: Strategies, research, and theory for college and university teachers.* (9th ed.). Lexington, MA: D. C. Heath.

Murray, H. G. (1980). *Evaluating university teaching: A review of research.* Toronto, Canada: Ontario Confederation of University Faculty Association.

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

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

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

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

Perry, R. P. (Ed.). (1990). Special Section: Instruction in higher education. *Journal of Educational Psychology, 82*, 183–274.

Sixbury, G. R., & Cashin, W. E. (1995a). *IDEA technical report no. 9: Description of database for the IDEA Diagnostic Form.* Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

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

Theall, M., & Franklin, J. (Eds.). (1990). *Student ratings of instruction: Issues for improving practice: New Directions for Teaching and Learning*, No. 43. San Francisco: Jossey-Bass.

IDEA Papers may be ordered from the Center. Individual copies are $2.00. A complete set of IDEA Papers may be ordered for $15.00. Bulk orders of the *same* paper: 20–49 copies are 30 cents a copy, 50–99 copies are 25 cents a copy, 100 or more copies are 20 cents a copy. Orders of less than $50.00 must be prepaid. Prices effective through 6/30/96.

Center for Faculty Evaluation and Development
Kansas State University
1615 Anderson Avenue
Manhattan, KS 66502–4073

ADDENDUM--IDEA PAPER NO. 32

Add the following as the last paragraph of the paper.

This paper has summarized the *general* conclusions from the research on student ratings. Whether those conclusions hold true for any given campus is an empirical question. If an institution has reason to believe that they do *not* apply, it should gather *local* data to answer the question. However, in the absence of evidence to the contrary, I suggest that the general conclusions serve as a guide.

10/95 WEC