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

IDEA Paper No. 22

January, 1990

Student Ratings of Teaching: Recommendations for Use

William E. Cashin
Kansas State University

" . . . 'statistics are no substitute for judgement.' One can also say that judgement is no substitute for statistics."
(Miller, 1987, p. 107)

This IDEA Paper compliments IDEA Paper No. 20 (Cashin, 1988) which summarized the research on student ratings of teaching and concluded that **student ratings tend to be reliable, valid, relatively unbiased, and useful.** If you accept those conclusions, then you will want to use that information in developing or revising your on-campus student rating system, or in selecting one that is commercially available. This paper attempts to derive recommendations based on that literature—a literature which, although sometimes based on *empirical* studies, is more often based on *experience* in using student ratings. (In some cases the recommendations may be based primarily on my *personal opinion*; those cases will be noted.)

The recommendations are divided into five sections: general considerations, the overall system, the student rating form itself, its administration, and its interpretation. In order to keep both the length of the text and the number of references manageable, I will presume that the reader is generally familiar with IDEA Paper No. 20. Therefore, I will usually *not* detail its conclusions and *nor* cite all of the references.

General Considerations

RECOMMENDATION 1—Use *multiple sources* of data about a faculty member's teaching if you are serious about accurately evaluating or improving teaching. The major writers (e.g., Centra, 1979; Miller, 1987; Seldin, 1980) all caution against using any single source of data. In IDEA Paper No. 21 (Cashin, 1989) I proposed an expanded definition of college teaching which argued that using only student rating data ignored several aspects of teaching.

RECOMMENDATION 2—Do use student rating data as *one source of data* about effective teaching, assuming you accept the conclusions of IDEA Paper No. 20.

RECOMMENDATION 3—Discuss and decide upon the *purpose(s)* that the student rating data will be used for *before* any student rating form is chosen or any data are collected. Before an individual (or an institution) develops or selects a student ratings system, one must first decide for what purpose or purposes the data will be used. The three most frequently mentioned purposes are described below.

EVALUATION—the data are used by faculty committees, academic administrators, etc., as *part of* the data upon which to base **personnel decisions: retention, promotion, tenure, or salary increases.** Every institution makes personnel decisions so evaluation is *necessarily* one purpose of any *institutional* student rating system.

IMPROVEMENT—student rating results are used by the instructor to make changes that he or she thinks will help the students learn more effectively or efficiently. It should be noted that "improvement" (or "development") does *not* necessarily imply a deficiency. Moving from a B-level to an A-level performance is definitely an improvement, but of an already strong performance. Although the institution's rhetoric often states that improvement is the primary purpose for using student ratings, frequently there is no systematic help provided by the institution for faculty whose student ratings suggest that improvement is needed.

ADVISING—the data are used by students and advisors to help in selecting instructors or courses. My impression is that relatively few institutions actually publish student rating data to help advise students.

All three of these purposes are *legitimate* uses of the data. However, not all student rating items serve every purpose equally well. Everyone involved—faculty, administrators, and students—should discuss and decide upon how the data will be used, i.e., what information will go to whom, *before* any ratings are collected. Such open discussion can do much to allay the legitimate concerns of the various parties involved and to enlist their cooperation.

The System

RECOMMENDATION 4—To obtain *reliable* student rating data, collect data from *at least ten raters* if this is possible. The average IDEA item reliability with ten raters is around .70. (All of the references to IDEA data are from Cashin & Perrin, 1978 unless otherwise noted.) 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. Data based on the ratings of fewer than

ten raters should be interpreted with *caution*, particularly if you want to generalize about what might be an effective way to teach the same course at a later date. However, combining data from several classes of less than ten students does yield reliable data.

RECOMMENDATION 5—To obtain *representative* student rating data, collect data from *at least two-thirds of the class*. This recommendation is based primarily on experience and common sense. Even using this guideline, one-third of the students would not be represented in the ratings. Some have suggested requiring ratings from three-quarters of the class, but experience with the IDEA system revealed that on average only about 70% of a class turn in ratings.

RECOMMENDATION 6—To *generalize* from student rating data to an instructor's overall teaching effectiveness, sample across *both* courses *and* across time. For improvement it is acceptable to look at the data from one course, but for evaluation you need a much broader sample (see Gillmore, Kane, & Naccarato, 1978). *I* suggest two or more different courses from at least three or more different terms.

RECOMMENDATION 7—For *improvement*, develop a student rating system that is *flexible*. Instructional goals vary widely from course to course, and so what is an effective method to teach one goal may not be effective in teaching another. Your student rating system needs to accommodate this diversity. **Cafeteria-type systems** provide the most flexibility. (See also Recommendation 20.)

RECOMMENDATION 8—Provide *comparative data*, preferably for all the items. Student ratings tend to be **inflated**. The average student rating on a 5-point scale is not 3.0—as one might think—but usually between 3.5 and 4.0. Also, ratings **vary widely** from item to item. On the 20 IDEA teaching method items, the lowest mean is 3.3; the highest, 4.3. Without comparative data it is *not* possible to meaningfully interpret student rating data.

RECOMMENDATION 9—Discuss and decide what *controls for bias* will be included in your system. Student ratings are correlated with variables *other* than the instructor's teaching effectiveness (Recommendations 10–13 will discuss specifics). The institution needs to decide what, if anything, will be done about these possible sources of bias.

RECOMMENDATION 10—Do *not* give undue weight to: the instructor's age, sex, teaching experience, personality, or research productivity; the student's age, sex, level (freshman, etc.), grade-point-average, or personality; or the class size or time of day when it was taught. These show *little or no correlation* student ratings. (See IDEA Paper 20 for references.) Regarding class size, although there is a tendency for smaller classes to receive higher ratings, it is a very weak inverse association, average r = –.09 (Feldman, 1984). The average correlation of class size for the 38 IDEA items is somewhat greater –.18 (Cashin & Slawson, 1977).

EXCEPTION, if the instructor provides evidence in his or her self-report for the influence of these variables, or if you or others have such evidence, that evidence should be taken into consideration.

RECOMMENDATION 11—Take into consideration the *students' motivation level* when interpreting student rating data. Student motivation tends to show higher correlations with other student rating items than any other variable. Instructors are more likely to obtain higher ratings in classes where students had a prior interest in the subject matter (Marsh, 1984), or were taking the course as an elective (Alemoni, 1981; Feldman, 1978). The average correlation of the IDEA motivation item, "I had a strong desire to take this course," with the other 37 items is .39.

RECOMMENDATION 12—Decide how you will treat student ratings from *different course levels*, e.g., freshman, graduate, etc. Higher level courses, especially graduate courses, tend to receive higher ratings (Alemoni, 1981; Braskamp et al., 1984). However, with the 38 IDEA items course level correlates only .07 on average.

RECOMMENDATION 13—Decide how you will treat student ratings from *different academic fields*. There is increasing evidence that different academic fields are rated differently (Braskamp et al., 1984; Cashin, Noma, & Hanna, 1987; Feldman, 1978; Marsh, 1984). What is not clear is why. For example, more quantitative courses—for example, math—tend to receive lower ratings. If you decide that this is because these courses are more difficult to teach, then you should take academic field into consideration when interpreting the data; if you think that certain fields are more poorly taught then you should *not*.

RECOMMENDATION 14—For *improvement*, develop a system that is *diagnostic*. The more diagnostic the system is, the more useful it will be for improvement. This means that the items included on the form should be **descriptive of specific and concrete teaching behaviors**. For example, "the instructor provided an outline for each class" is more specific than an item like "the instructor gave clear presentations." (See also Recommendation 19.)

RECOMMENDATION 15—Develop a system that is *interpretable*. It is very important that the data be *understandable* to the average faculty member. **Using words as well as numbers** is one way to achieve this. **Including a written explanation** along with the results is also desirable—although experience suggests that many faculty will *not* read it. The ideal solution is to **have one or more faculty consultants on your campus** who are available *both* to help faculty understand their ratings *and* to suggest ways that they might improve their teaching if that is appropriate.

The Form

RECOMMENDATION 16—For evaluation, *use a few global or summary items or scores*. This recommendation is more a *personal opinion* but such summary, or global, student rating items tend to correlate more highly with student learning than do more specific items (Cohen, 1981).

Suggested summary items are:
1) Overall, how effective was the **instructor**?
2) Overall, how worthwhile was the **course**?
3) Overall, how much did you **learn**?

The students' ratings on these items would be **like a final course grade** in that the instructor would have some idea of how the students rated him or her, but would not know why. However, such items would serve the purpose of evaluation which is to decide how well the instructor taught (*not* what he or she might do to improve—which is the focus of development). Using a form with only a few items has some distinct advantages. Such items apply to a wide variety of courses, (probably to all courses) and so can be used as the basis of comparison *across the institution*, **as long as the appropriate comparative data are available**. Using such a short form also avoids wasting the students' time and the institution's money.

RECOMMENDATION 17—Use the short, evaluation form (or items) in every class every term. Using such a form can flag courses that may be ineffectively taught—so that more extensive data can be collected next term—but it avoids using a long, diagnostic form in classes which historically have received acceptable ratings.

RECOMMENDATION 18—Use a *long, diagnostic* form *in only one course per term*—in the course that the instructor wishes to focus upon for improvement. Most instructors would be doing well to **improve one course a term**. Using a diagnostic form in only one course a term **focuses the instructor's efforts** and avoids gathering data that may not be used.

RECOMMENDATION 19—For improvement, use *items that require as little inference as possible* on the part of the student rater and as little interpretation as possible on the part of the instructor. This is a corollary of Recommendation 14 that improvement systems need to be diagnostic. Concrete items descriptive of specific behaviors tend to be most helpful to an instructor looking for suggestions about how to improve.

RECOMMENDATION 20—For improvement, do *not* use a *single, standard set* of items for every class. *Provide a pool of items* or *some kind of weighting system*. This is a corollary of Recommendation 7 on flexibility. The problem with using a form which contains a single set of items is that it assumes that there is a *single, correct way to teach*, and that *every instructor in every class* should do all of the things listed on the form. Different course objectives—and probably different student learning styles—require different methods. One solution to the flexibility problem is to use a pool of items as the cafeteria systems do. The instructor selects only items that fit his or her course for the students to rate. IDEA uses a weighting system where the instructor, a faculty committee, etc. weight how important a given item—in IDEA's case general course objectives—is for the given course. Teaching methods are flagged for the instructor's consideration *only* if the research shows that the method is *relevant* to the goals selected for that course.

RECOMMENDATION 21—Use a *5-point to 7-point scale*. Scales with less than "5" points do not discriminate as well, but using more than "7" points adds little. (There are a number of other technical considerations discussed in the literature, but little consensus on what is best. Interested readers can consult Berk, 1979; or Doyle, 1983.)

RECOMMENDATION 22—In the analysis of the results, report computations *only to the first decimal place*. Although primarily a *personal opinion*, even reporting data to only the first decimal place yields 41 points on a 5-point scale (1.0 to 5.0). Most student rating data—as most of our classroom exam data—are *not* that precise, i.e., a 4.0 is rarely different from a 3.9.

RECOMMENDATION 23—Do *not overinterpret the data*, allow for a margin of error. This is a corollary of Recommendation 22. Depending upon the standard error of measurement of the items, scores within + or –.3 or more may not really be different. Combining the data into a limited number of categories, *perhaps ten*, rather than using all 41 points is both more understandable and more realistically reflects the level of accuracy of the data.

RECOMMENDATION 24—Use *frequency distributions*—what number or percent of the students rated the item "1" or "2," etc. These are more understandable to most faculty than calculating a standard deviation for each item. Also, the distributions can contain useful information. If all of your ratings are high, keep doing whatever you do. If they are all low, stop. But what if the distribution tends to be *flat*, the raters tended to pick all of the numbers equally; or what if the distribution tends to be *bimodal*, the ratings cluster at the two ends? The latter may mean that what you are doing works for one group of students but not for another. You probably need to keep doing what you are doing for the one group, and add something new for the other. First, you will have to figure out who the two groups are. The most common groupings are majors and non-majors.

RECOMMENDATION 25—For *improvement*, ask for *open-ended comments* as well as quantitative ratings. Sample items are:

1) Describe one or more things about the course that you **found helpful**,
2) What suggestions do you have about **how the course might be improved**.

The comments which the students make in responding to these kinds of questions can be particularly helpful for improvement. Often these comments will help explain why you received low ratings on one of the quantitative items. They can also provide suggestions about some changes you might make to help the students learn better. I would *not* substitute open-ended questions for quantitative ones, however. The two types *compliment* each other. Sometimes just reading the students comments gives a negative impression while looking a the numerical ratings shows relatively high numbers.

RECOMMENDATION 26—Use the open-ended comments *only* for improvement. *My* reasoning for this recommendation is that, especially for promotion and tenure decisions, there can be hundreds or even thousands of comments. To assess them accurately one should do a content analysis classifying every response as to content and also making a judgment about how positive or negative the comment is. This is *extremely time consuming*. My belief is that usually only the individual instructor has the motivation to do this and so the comments should only be used by the instructor for improvement. To have evaluators simply scan the comments to gain a general impression opens up the possibility that what will be remembered will be the more sensational comments, not the more representative ones. The literature is split on this question, however. Other writers recommend using these comments for evaluation as well as for improvement. Some say include *all* of the comments—which leads to a time problem. Other writers suggest including only a *random sample* of comments—in small classes especially, this could lead to basing personnel decisions on a small, possibly unrepresentative sample.

Administration

RECOMMENDATION 27—For evaluation, *develop standardized procedures* covering all relevant aspects of your student rating system and *monitor that the procedures are followed.* When the student rating data are going to be used for personnel decisions, it is important that everyone have confidence that everyone's data were gathered and treated fairly. Recommendations 28–32 deal with some specifics.

RECOMMENDATION 28—For evaluation, *administer the ratings about the second to the last week of the term.* You want to obtain the ratings near the end of the term so that the students will have an accurate perception of the *total* course and of what they have learned, but you do *not* want to give them so close to the end of the term that the students will be distracted by concerns about getting assignments in, or about what will be on the final exam. Avoid administering the ratings on the last day of class or on the day of the final exam. (See Lowman, 1984 for a persuasive rationale for administering the student ratings on the last day of class *if* you use the last class as he suggests.)

RECOMMENDATION 29—Develop *standardized instructions* that include the purpose(s) for which the data will be used, and who will receive what information, and when. Ideally, these instructions will be printed on the top of the student's form and also *read aloud by the proctor.*

RECOMMENDATION 30—Instruct the students *not to sign their ratings.* Studies suggest that requiring signatures will inflate the ratings (Braskamp, et al., 1984; Feldman, 1979; Marsh, 1984). Also, tell the students *if the instructor will be given their handwritten responses.* By doing this, the students can print or leave the open-ended questions blank if they are concerned about confidentiality and possible retaliation from the instructor. Some instructors have the open-ended questions printed on a separate sheet of paper which they let the students take home so they can type their comments if they wish. It also allows the students to give more thought to their responses. If we want the students to cooperate by giving us their honest feedback, we must do everything we can to insure confidentiality.

RECOMMENDATION 31—The instructor may hand out the rating forms and read the standardized instructions, but the instructor should *leave the room until the students have completed the ratings and they are collected.* When the instructor remains in the room, the ratings tend to be higher (Feldman, 1979; Marsh, 1984).

RECOMMENDATION 32—The ratings should be collected by a *neutral party* and the data taken to a predetermined location—often to where they are to be scored—and they should *not* be available to the instructor until the grades are turned in. This is the conventional wisdom with which I strongly agree. Following these procedures does much to insure people's confidence in the fairness of the system.

Interpretation

RECOMMENDATION 33—Develop a *written explanation* of how the analyses of the student ratings are to be interpreted. For data from individual classes that will be interpreted in isolation—particularly if being used for evaluation—include the caveats about having at least ten raters and about having data from at least two-thirds of the class. Remind readers that if they want to generalize to the instructor's overall teaching effectiveness they should have ratings from two or more courses from three or more terms. Above all, remember that student ratings are only one source of data.

RECOMMENDATION 34—Appoint a faculty member to serve as *instructional consultant* to help faculty interpret their results and to improve their teaching (Cohen, 1980). Give this consultant release time to fulfill these consulting responsibilities and support to develop the desirable skills. (There is a considerable literature on instructional improvement, some of which has been described in other IDEA Papers.)

References

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

Berk, R. A. (1979). The construction of rating instruments for faculty evaluation. *Journal of Higher Education, 50,* 650–669.

Braskamp, L. A., Brandenburg, D. C., & Ory, J. C. (1984). *Evaluating teaching effectiveness: A practical guide.* Beverly Hills, CA: Sage.

Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research.* IDEA Paper No. 20. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E. (1989). *Defining and evaluating college teaching.* IDEA Paper No. 21. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E., Noma, A., & Hanna, G. S. (1987). *IDEA technical report no. 6: Comparative data by academic field.* Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E., & Perrin, B. M. (1978). *IDEA technical report no. 4: Description of IDEA Standard Form data base.* Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E., & Slawson, H. M. (1977). *IDEA technical report no. 2: Description of data base 1976–1977.* Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Centra, J. A. (1979). *Determining faculty effectiveness: Assessing teaching, research, and service for personnel decisions and improvement.* San Francisco, CA: 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.

Doyle, K. O. (1983). *Evaluating teaching.* Lexington, MA: D. C. Heath.

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' rating of their teachers and courses. *Research in Higher Education, 10,* 149–172.

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

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.

Lowman, J. (1984). *Mastering the techniques of teaching*. San Francisco: Jossey–Bass.

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

Miller, R. I. (1987). *Evaluating faculty for promotion and tenure*. San Francisco, CA: Jossey–Bass.

Seldin, P. (1980). *Successful faculty evaluation programs: A practical guide to improve faculty performance and promotion/tenure decisions*. Crugers, NY: Coventry Press.