# Design and Use of Student Evaluation Instruments in Instructional Development

**Richard A. Schwier**
*Associate Professor of*
*Educational Communications*
*College of Education*
*University of Saskatchewan*
*Saskatoon, Saskatchewan S7N 0W0*

**Abstract.** What can student evaluations of instruction contribute to formative evaluation? Five decades of research indicate that student evaluations can be developed which are reliable, and which can reflect student satisfaction. This article discusses the roles played by student evaluations in the instructional development process, reviews issues related to their use, and outlines suggestions for the construction, administration and analysis of student evaluations by the instructional development consultant.

Well-designed instruction can fail as a result of weak teaching performance, and relatively weak instruction can be judged as palatable based upon the performance of a strong teacher. While assessments of instruction can include a variety of strategies such as expert observation, analysis of student achievement data, self-appraisal, videotaping, and peer evaluation, one of the most important sources of diagnostic information has been student evaluations. Although student evaluations are used commonly, they have been subject to controversy. This discussion addresses issues surrounding student evaluations of instruction within the context of instructional development programs and reviews suggestions for the development and utilization of diagnostic instrumentation. The issue of whether or not student evaluations of instruction should be used will not be addresssed, although it is noted that some individuals argue that student evaluations are hopelessly flawed. Given the fact that student evaluations are used extensively, the focus here is to explore ways to minimize contamination of these data.

### Roles Played by Student Evaluations of Instruction

Student evaluations of instruction are used for three purposes: for promotion and tenure decisions, for diagnosis and remediation of teaching skills, and for student use in selecting courses and instructors (Derry, Seibert, Starry, Van Horn, & Wright, 1973; Kulik, 1976; Sheehan, 1975; Shingles, 1977; University of Alberta, 1979). These purposes do not describe fully the roles that student evaluations of instruction have played in the process of instructional development. How can student evaluations be used in a developmental context?

*Placebo.* Student data can be gathered and ignored. Although fruitless effort and mindless activity are not generally a good idea, just the act of gathering information can give the client a sense that progress is being made. Obviously, this is not recommended practice in a development context, as it can lead to credibility problems later in a project. Nevertheless, in practice, this is one activity that a client often expects, and meeting the expectations of the client early in the development process can enhance the developer-client relationship.

*Ice-breaker.* After a relationship between a developer and a client is established, concerns often surface which include: Do the students like the class? Are they bored with this? Do they like me? How am I coming across? Of course these concerns are not always voiced, but it is surely an insensitive consultant who has not felt these questions rumbling just beneath the surface of a conversation. Until these concerns are addressed, meaningful progress may be impeded in other aspects of the development process. Student evaluation data can be used as a reason to address these issues.

*Product Appraisal.* Student opinions of the instructional product can be acquired through student evaluations. Field testing a product requires sampling the attitudes of those affected, dispassionate or not. Comparing these opinions to other sources of information can often provide insights which are overlooked by persons close to the development process, or overly concerned about content. With due caution about the validity of any type of opinion-sampling procedure, student impressions of the difficulty, sequence, entertainment value, and approach of the task can be useful to the developer.

*Instructional appraisal.* Student evaluations can be used to identify perceived instructor weaknesses and strengths, and thus be used along with other data sources in a diagnostic role. The developer should handle this information skillfully, or as much harm as good can result. Unless the client is open to criticism, it is better to deemphasize the role of student evaluations in instructor appraisal, and focus on the appraisal of the instructional product or task.

Given the variety of ways in which student evaluations of instruction can be used in a development context, it is useful to examine some of the issues surrounding their construction and interpretation, and then discuss some strategies which can be used to minimize contamination of these data.

## Issues and Variables

There are a number of important issues related to the construction and validation of student evaluations of instruction. A discussion of major issues follows, and each will be preceded by a summary statement, which briefly provides the author's interpretation of available evidence.

### Reliability

*Student evaluation instruments can be tested for reliability.* One of the arguments most often leveled at student ratings of instruction is that such ratings are unstable. This argument

contends that instructor ratings are affected by contaminating variables such as immediate experiences in the class and the characteristics of instruments. Testing five core items used university-wide as part of a large diagnostic approach to instructor ratings, Seibert (1977) reported inter-item correlations ranging between .70 and .85. Seibert pointed out that the high and consistent intercorrelations presented certain problems, however. The high correlations could be interpreted as lower bound estimates of reliability, and high reliability items could be considered stable. But the high intercorrelations also indicated a great deal of shared variance between the items. Thus, the items were not measuring mutually exclusive characteristics; they were not independent. Nevertheless, the findings were consistent with earlier studies which found high correlations between global ratings and scores on the rating instruments (Cohen & Humphries, 1960; Harvey & Barker, 1970), even-versus-odd numbered items (Lovell & Haner, 1955), and randomly paired items (Maslow & Zimmerman, 1956).

As indicators of reliability, these results are favorable, but the problems associated with high item intercorrelations must be underlined. Most student evaluations of instruction, particularly if the intended purpose is diagnostic, are meant to measure several different dimensions of teaching. The reported studies, on the other hand, indicated that instruments used were largely unidimensional.

But in general, the span of years covered by these studies, the consistently high correlations, and the wide variety of instruments, methodologies and subjects studied, suggest that reliability is not a fleeting phenomenon in instructor rating instruments. Instrumentation can be developed which satisfies reasonable reliability standards.

### Validity

*Instruments appear to reflect student satisfaction rather than teacher effectiveness.* Do student evaluations of instruction measure instructional effectiveness? This critical question has been studied, hotly debated, and occasionally disregarded, but it must be answered before interpreting data.

One can argue that evaluation instruments exhibit face validity (appear to measure effective teaching characteristics—logical appeal of items). General agreement was found on the characteristics of effective teaching (Hildebrand, Wilson, & Dienst, 1971). Student surveys (Sheffield, 1974) and a review of teacher-rating instruments (Wotruba & Wright, 1975) identified similar criteria, and this author noted considerable item redundancy in a wide sampling of instruments used by different universities in North America for student evaluation of instruction. While this could certainly be the result of instrument inbreeding, it can be argued that face validity exists between student rating instruments and the opinions of professionals and students. Many validity issues can be dealt with through a "cafeteria approach," which is a system used by faculty to create individualized instruments (a more detailed description of this approach follows later).

A popular method of determining validity has been to correlate student evaluations of instructors and student achievement. The obvious rationale is that students learn more from the more effective teachers. This line of study has exihibited its own validity problems, as affective areas of growth have been ignored in favor of cognitive performance measures. Nevertheless, several studies isolated mild, but consistently positive correlations between instructor evaluations and student achievement (Bryson, 1974; McKeachie, Lin, & Mann, 1971; McKeachie & Solomon, 1958; Meinkoth, 1971; Morsh, Burgess & Smith, 1956; Sullivan & Skanes, 1974). Only Rodin and Rodin (1972) found a negative relationship. The research suggests a small, but significant relationship between student achievement and instructor ratings.

There are several other factors which contaminate the validity of instructor rating instruments (Sheehan, 1975). First, it is difficult to establish norms for the purpose of comparison. While difficult, an existing project (Seibert, 1977) has had considerable success establishing university-wide norms for a pool of 200 evaluation items. Second, rating scales assume that all items are of equal importance, and as this is not usually the case, interpretation of resultant data is invalid. This legitimate criticism can be dealt with through the use of a dual rating scale, one rating instructor performance on the item, and another rating item importance. Finally, influence-peddling tactics by the instructor during the administration of evaluations can contaminate results. This appears to be a significant factor, and certainly guidelines for the unbiased administration of evaluations are essential.

In a recent report, the University of Alberta (1979) concluded that student evaluations describe teaching accurately, but seem to measure student satisfaction rather than teacher effectiveness. Student-based appraisals are a seductive source of information. Clients often give student evaluation data more credence than they should, even though there are serious validity problems with this source of information. The developer, as moderator between the possible misconceptions of a client and questionable data, must be wary of relying too heavily on student evaluations to assess instructional effectiveness. But as one of a complement of measures, student evaluations of instruction can provide useful information about the attitudes of the consumers of instruction.

### Class Size

*Small and large classes may rate the instructor higher than other size classes.* A myth has developed that teachers with smaller classes receive more favorable ratings by students than teachers with larger classes. Generally, this is not supported by the literature (Costin, Greenough, & Menges, 1971), but Gage (1961) found an interesting interaction between class size and instructor rating such that small and large classes were rated higher than classes with 30-39 students. A recent study (Morsh, Overall, & Kelser, 1979) supported this interesting finding, suggesting that the relationship between class size and evaluations by students is nonlinear. Large and small classes were rated more favorably than other classes, but the magnitude of the effect varied across different components of the evaluation. "Quality of group interaction" accounted for a significant amount of variance, while other components were relatively weak.

### Compulsory Instruction

*Compulsory classes may result in lower instruction ratings.* Perhaps because of negative attitudes of the students toward a required course, and perhaps because of negative instructor attitudes toward having to teach compulsory classes, there is a reason to suspect that the ratings of instructors are affected by this variable. Cohen and Humphries (1960), Gage (1961), Lovell and Haner (1955), and Pohlmann (1975) found that students required to take a course or students in required courses rated instructors

lower than students electing to take courses. An earlier study by Heilman and Armentrout (1936) did not find this relationship, but the evidence is sufficiently persuasive to conclude that compulsory classes are a significant factor, and should be taken into consideration in the interpretation of data for instructor evaluation.

### Personality Characteristics

*Teachers who exhibit warmth and culture may receive higher ratings.* There is considerable evidence to suggest that personality variables, as well as teaching skill, can affect student ratings of instruction. A review of several studies suggested that teacher warmth contributes to perceived teacher effectiveness (University of Alberta, 1979), and Isaacson, McKeachie, & Milholland (1963) concluded that teachers who were cultured (artistic, polished, imaginative, effectively intelligent) were rated favorably on measures of teaching effectiveness by students.

The "Dr. Fox lectures" provided an interesting investigation of the impact of charisma on the evaluation of an instructor (Naftulin, Ware, & Donnelly, 1973). In this study, an actor lectured professional educators on the topic of "the application of mathematics to human behavior"—a bogus topic. The actor taught charismatically, but presented irrelevant, meaningless, and conflicting content. Despite the nonsubstantive nature of the lecture and discussion, the presentation was rated favorably by a wide majority of participants. Costin, Greenough, and Menges (1971) challenged earlier conclusions, stating that design and conceptual flaws in existing studies minimized their worth; however, they agreed that the influence of personality factors on student evaluations of instruction was intuitively persuasive, and that further study was needed in this area, taking both student and teacher personality factors into consideration. Indeed, studying interactions between teacher and student personality factors and instructional methods represents a challenging area of investigation.

### Gender

*There is no significant difference between the ratings of men and women teachers, or between ratings made by male and female students.* Most of the studies of student evaluations of instruction were conducted prior to 1970, and relatively few of the studies explored interactions between the sex of the evaluator and the instructor on ratings. Bendig (1953), Caffrey (1969), Downie (1952), Lovell and Haner (1955), and Remmers (1939) reported no significant difference between the sexes on overall ratings. More recent studies reported tendencies for female students to favor instructors who were rated high on skill and structure dimensions (McKeachie, Lin, & Mann, 1971), and for female students to favor female instructors (Walker, 1969). But all things considered, sex does not represent a significant variable in the construction of a strategy for student evaluations of instruction.

### Course Grades

*Actual or anticipated grades do not appear to affect instructor ratings.* There is considerable concern expressed that student ratings are affected by the grade received or expected in a course. This line of thinking follows to the logical conclusion that teachers who give "high" grades will be better liked by their students, and will therefore receive higher ratings. Clearly, the literature does not support this contention. No relationship between grades and ratings was found in a number of studies (Cohen & Humphries, 1960; Eckert, 1950; Guthrie, 1949, 1954; Voeks & French, 1960), and several other studies isolated only weak relationships (Caffrey, 1969; Rayder, 1968; Stewart & Malpass, 1966; Treffinger & Feldhusen, 1970; Walker, 1969). Even if the reported correlations were not spurious, the inconsistency of findings suggests that grades are NOT a significant factor.

### Instructional Improvement

*Student evaluations alone do not generally produce instructional improvement.* Student evaluations of instruction have not been shown to produce changes in instructional behavior, generally (Centra, 1973; Pambookian, 1974). These studies suggested that this was not true of all instructors, however. Greatest change was noticed in instructors who had rated themselves higher than their students rated them. The need for assistance in the interpretation of evaluations and the improvement of teaching skills was also indicated. It is not sufficient to evaluate instruction; mechanisms must exist for the improvement of teaching behavior based upon the evaluations.

## Design of Evaluation Instruments

Given this background on the use of student evaluations of instruction, what should be included in an effective evaluation instrument? What characteristics should be evaluated, how should items be written, and what structural characteristics of the instrument are important?

### Selecting Course and Instructor Characteristics

The number of items which have been developed to test characteristics of instruction is staggering, and many have not been subjected to any form of analysis. Commonly, a developer or instructor has decided what is important in a particular course and devised a series of questions to assess these concerns. This has resulted in a plethora of scales of unknown reliability and validity, which may or may not assess the range of characteristics which affect the instructional process. There is also a smaller yet comprehensive collection of items reported in the literature which has been subjected to rigorous testing, and which displays reasonable reliability (e.g., Hildebrand, Wilson, and Dienst, 1971).

There appears to be general agreement on the characteristics of effective instruction, and the instructional developer should take them into consideration when writing evaluation items. These characteristics fall under three major headings: instructor behavior, learner outcomes, and course components (see Table 1). Items related to these three characteristics should be clearly separated for diagnostic purposes.

### Writing Stems

Scissons (1979) outlines rules for the development of items or statements used for evaluation, sometimes called "stems."

(1) State stems clearly—they should be specific and objective.

(2) Measure only one trait, behavior, or activity with each item (e.g., a single item should not be used to evaluate an instructor's warmth and knowledge of subject matter).

(3) Evaluate only what has happened or is happening. Prediction of future action is unreliable at best.

Examples of well-written and poorly-written stems are presented in Figure 1. In this case the poorly-written stem attempts to measure two traits instead of one.

Table 1.

Categorization of Characteristics of Effective Instruction.*

| Instructor Behaviors | Learner Outcomes | Course Components |
|---|---|---|
| knowledge of subject matter | knowledge and skills | course aplicability |
| enthusiasm, dynamism | interests and curiosity | written assignments |
| communication skill | self concept | reading assignments |
| difficulty, loading, pace | social skills and attitudes | textbook |
| organization, clarity | vocational skills and attitudes | examinations |
| feedback | | grading |
| interaction | | media |
| rapport | | teaching assistant |
| flexibility | | laboratory assignments |
| | | recitation section |

*Characteristics drawn from Hildebrand (1971), Kulik (1976), Scissons (1979), Seibert (1979), Sheffield (1974), and Wotruba & Wright (1975).

| | None of the Objectives | Some | Most | All Objectives |
|---|---|---|---|---|
| *Well-Written* | | | | |
| Examinations tested the objectives outlined for the course | 1 | 2 | 3 | 4 |
| | Low | | | High |
| *Poorly-Written* | | | | |
| Examinations were good, and the instructor seemed open to criticism | 1 | 2 | 3 | 4 |

**Figure 1. Examples of well-written and poorly-written stems.**

| | never | | | | always |
|---|---|---|---|---|---|
| 1. The instructor summarized units of content | 1 | 2 | 3 | 4 | 5 |
| | never | 1-2 times | 3-4 times | 4-5 times | over 5 times |
| 2. The instructor summarized units of content | 1 | 2 | 3 | 4 | 5 |

**Figure 2. Examples of different usages of anchor points for the same stem.**

### Writing Anchor Points

Anchor points are the descriptors used to evaluate the individual on each stem. Typically, anchor points describe a five-point continuum from "excellent" to "unsatisfactory," or "always" to "never." While such designations are in common use, they do not always exhibit the precision and relevance necessary to adequately describe the characteristic being evaluated.

There are five issues to consider when writing anchor points (Scissons, 1979).

(1) Clarity. Use short, simple terms which are easily understood by the intended audience.

(2) Consistency. Terms should be consistent with the stem.

(3) Definition. Define each point on the scale with a distinct rank position in order to increase precision. The user should not have to guess at the difference between a "three" and a "four" on a five-point scale.

(4) Variety. Use diverse cues or language in defining points in order to add variety and precision.

(5) Objectivity. Use objective terms. Moral, ethical, or social considerations, or terms such as "good" or "unacceptable" actually provide little useful data.

Two examples of anchor points using identical stems are presented in Figure 2. Of course this represents an extreme example, but it illustrates the point that careful use of anchor points can provide the user with more information than carelessly applied anchor points.

The instructional developer must not only ensure that items exhibit sound stem and anchor point construction, but also use discrimination analysis on the item pool. Hildebrand, Wilson, and Dienst (1971) analyzed a large number of items in terms of their ability to discriminate between "good" and "poor" teachers. This was the only study encountered which utilized discrimination analysis, and it appears that this is a desirable procedure in the analysis of items for adoption. Obviously, items which do not discriminate well are not useful for evaluation, although limited diagnostic purposes could be served by weak items.

### Using Dual Rating Systems

One of the primary causes of invalidity in student evaluations of instruction is the underlying assumption that all of the items are of equal

| Descriptors | Frequency of Occurence | | | | | Importance to the Course | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | never | seldom | sometimes | often | always | unimportant | slight importance | important | very important | extremely important |
| The instructor: | | | | | | | | | | |
| 1. gives organized lectures | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| 2. motivates the class | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |

**Figure 3. Dual rating scale items.**

importance (Sheehan, 1975). One way to circumvent this is for students to select important teaching behaviors and then rate instructors. In this way, the instructor would learn which behaviors are most important to the students, and ratings would be performed by students who perceive the listed behaviors as important. An alternative approach omits having the students select teaching behaviors. Called the dual or double rating scale, this type of instrumentation allows the students to rate each item as to its perceived importance and independently evaluate the instructor on that item. In this way, an instructor rated very low on two items could select the most important instructional component on which to concentrate improvement resources. Figure 3 provides an example of two items from a dual rating scale.

Utilizing this approach allows the developer and client to evaluate emphasis as well as competence in a number of areas. It is important to note that this will provide a measure of *perceived* importance. It is possible that the judgements of professionals and students will differ, and consequently, data must be subjected to careful interpretation. (For an interesting discussion of methods of analyzing similar data types, see Misanchuk, 1981.)

### Adopting a Cafeteria Approach

One of the most flexible approaches to student evaluations encountered to date is called the "cafeteria" system. Developed by the Measurement and Research Center at Purdue University, the cafeteria approach to instructor evaluation is not a single instrument, but rather a set of procedures which are used by faculty to create individualized instruments. Briefly, a large pool of statements describing desirable teaching behavior is stored centrally on magnetic disks, and a printed catalogue of the items is distributed to the faculty. Using this data base, each instructor can select items which are relevant to specific needs, and the central computer prints out a customized survey form.

The flexibility offered by this system is attractive, and addresses many of the problems faced in the development of standardized evaluation strategies. The problem of increasing volume and heterogeneity of instructor and course evaluations is minimized by centralizing the process and standardizing items. Normative data are collected over time to provide reliability and validity measures and can be used for comparative purposes. The system allows a standardized yet flexible schema for assessing non-parallel teaching situations.

The cafeteria approach does present a problem which has not been addressed in the literature. As instructors are responsible for selecting the majority of the items, there could be a tendency for them to select items which make them look as good as possible. Most instructors are aware of their own strengths and weaknesses, and this system provides the opportunity to exploit those items which tests strengths. There is a danger that resultant data would reflect the instructor's ability to construct a clever questionnaire. In the same manner, an instructional developer might be tempted to gather biased data which would enhance the image of the instructional product or client.

## Conducting Evaluation

Since student evaluations are surveys, care must also be taken in standardizing administration procedures to reduce the effect of instructor-imposed bias. Suggested procedures include: (1)Having the instructor leave the room while evaluation instruments a administered, (2) Making a student responsible for administration, and (3) Delaying reporting of results to instructors until grades have been recorded and filed with the university.

## Procedures for Analyzing Data

It is a relatively easy job to summarize student responses, but without specific mechanisms for making decisions based upon results, the data become useless. The needs of the developer and client will dictate the analytical approach employed, but generally speaking, either norm-free or norm-referenced analysis will be used.

### Norm-Free Analysis

Analyzing data without reference to established norms assumes that the instructor has established personal criteria for making decisions. Overwhelming positive or negative

responses from the students clearly indicate success or weakness, but difficulties arise when responses are ambivalent.

How much credence should be given to a set of responses that is positively skewed, but widely distributed? This may depend upon the relative importance of the item or item cluster to the particular instructor. If the instructor feels that a particular aspect of teaching is important, then a very favorable response would be required to indicate success. Ambivalent response patterns to an item of less importance may not indicate an area of special concern. Often, the median response is used to determine success or failure, and while it is helpful and easy, this type of interpretation is based upon only limited information and could lead to faulty conclusions. In all, norm-free analysis can be useful, but analysis should be based upon predetermined criteria whether self-selected or imposed.

**Norm-Referenced Analysis**

Although the development of normative data would take a great amount of time, there are a number of advantages to norm-referenced analysis. First, most teachers obtain ratings that are above the middle ranking point for items. In other words, most teachers will average above 3.0 on a five-point scale on many items. Comparing an individual's score on a specific item with the scores of everyone using the item will give that individual a more accurate indication of relative standing. This can provide valuable information on the difficulty of objectives and shed light on real or imaginary successes and failures.

Generally, percentile rankings or blocked means by percentile can be used to roughly indicate the instructor's position on any item, relative to other instructors in a defined population. Kulik (1976) cautioned against the overinterpretation of these types of results. It must be remembered that using percentile rankings automatically relegates half of the users into "below average" categories. If every instructional program field-tested exhibited exemplary motivational approaches, half would still be considered sub-standard.

The developer and client can also benefit from establishing personal norms over a number of years. Areas of concern can be identified, and personal growth measured using similar evaluations over an extended period of time.

## Managing Client Relationships

As a source of formative data, the role to be played by the student evaluation component will dictate the emphasis placed on results. Student evaluations of instruction, serving as either a placebo or an ice-breaker, will lend little, if any, formative information to the instructional development process. Rather, they will merely serve to placate the client. Considerably greater attention will be placed on these data, however, when the purpose is to appraise the instructional product or the instructor. In these roles, the aforementioned suggestions for construction and administration assume greater signficance in the process of instructional development.

In addition, given the volatile nature of this source of data, additional cautions are necessary if the integrity of the overall evaluation strategy is to survive.

*Use student evaluations only with the informed consent of the client.* Of course this principle should be true of any activity an instructional developer carries out on behalf of a client, but it is particularly important for student evaluations. Be sure the client clearly understands the purpose of the exercise, encourages its use, and understands the limitations of this type of information.

*Jointly construct or adopt evaluation instruments.* Ask the client to participate in the development of evaluation instrumentation. This will not only promote the growth of a healthy professional relationship, it may also help the client and developer articulate concerns. If joint construction is not possible, then you should at least review adopted instruments with the client prior to their administration.

*Administer the instruments and summarize the data collected prior to meeting with the client to discuss results.* Before meeting with the client, put the data in a form which will promote efficient review. Do this as soon as possible following the administration of the evaluations, because in most cases the client will be anxious to see the results.

*Jointly review and discuss results to avoid overinterpretation or misinterpretation of data.* Because of personal involvement, the client may focus on certain aspects of the evaluation and miss others. It is your responsibility to bring an objective point of view to the discussion, and to prepare yourself for the meeting by reviewing the information and anticipating client reactions.

*Compare the results with other data.* Because of the questionable reliability and validity of student evaluations of instruction, it would be unwise to allow student evaluations to represent the only major component in your data collection strategy. Use other data sources, such as expert observation, peer appraisal, and self evaluation, and focus the discussion on data from all of these sources.

*Wherever possible, emphasize the positive results of the evaluation.* This is not to suggest that you ignore negative feedback; however, you will probably spend a substantial amount of time improving weaknesses as the development process continues, so a positive approach during diagnosis will help promote an open relationship. When negative issues must be confronted, be direct and honest. In other words, deal with positive results first, but do not use them to avoid more difficult discussion.

*Be prepared to offer alternative strategies to remediate identified deficiencies.* Because you have an opportunity to review the data prior to meeting with the client, areas of concern can be identified and remediation strategies outlined before discussion. This will not only lead to a streamlined meeting, but will also help to alleviate any client anxieties which might arise from negative student feedback. Weaknesses will not seem as threatening if ideas for improvement can be identified immediately.

In conclusion, student evaluations, judiciously employed, can make a contribution to a comprehensive evaluation package. They are a fact of life in most multi-component evaluation strategies, and given careful attention to their construction, administration, and interpretation, they represent an important factor in the instructional development process.

## References*


Bendig, A. W. Student achievement in introductory psychology and student ratings of the competence and empathy of their instructors. *Journal of Psychology*, 1953, *36*, 427-433.

Bryson, R. Teacher evaluations and student learning: A reexamination. *Journal of Educational Research*, 1974, *68*(1), 12-14.

Caffrey, B. Lack of bias in student evaluations of teachers. *Proceedings of the 77th Annual Convention of the American Psychological Association*, 1969, *4*, 641-642.

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

Centra, J.A. The student as godfather? The im-

pact of student ratings on academia. *Educational Researcher*, 1973, 2, 4-8.
Centra, J. A. The relationship between student and alumni ratings of teachers. *Educational and Psychological Measurement*, 1974, *34*, 321-325.
Centra, J. A., & Creech, F. *The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness.* Princeton, N.J.: Educational Testing Service, 1976.
Cohen, J., & Humphries, L. G. Memorandum to faculty. University of Illinois, Department of Psychology, 1960. (mimeographed)
Costin, F. A graduate course in the teaching of psychology: Description and Evaluation. *Journal of Teacher Education*, 1968, *19*, 425-432.
Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, 1971, *41*(5), 511-535.
Costin, F., & Grush, J. Personality correlates of teacher-student behavior in the college classroom. *Journal of Educational Psychology*, 1973, *65*(1), 35-44.
Derry, J. O., Seibert, W. F., Starry, A. R., Van Horn, J. W., & Wright, G. L. The cafeteria system: A new approach to course and instructor evaluation. Purdue University: Measurement and Research Center, 1973. (mimeographed)
Downie, N. W. Student evaluation of faculty. *Journal of Higher Education*, 1952, *23*, 495-496.
Eckert, R. E. Ways of evaluating college teaching. *School and Society*, 1950, *71*, 65-69.
Gage, N. L. The appraisal of college teaching. *Journal of Higher Education*, 1961, *32*, 17-22.
Guthrie, E. R. The evaluation of teaching. *Educational Record*, 1949, *30*, 109-115.
Guthrie, E. R. *The evaluation of teaching: A progress report.* Seattle, Washington, University of Washington, 1954.
Harvey, J. N., & Barker, D. G. Student evaluation of teaching effectiveness. *Improving College and University Teaching*, 1970, *18*, 275-278.
Heilman, J. D., & Armentrout, W. D. The rating of college teachers on ten traits by their students. *Journal of Educational Psychology*, 1936, *27*, 197-216.
Hildebrand, M., Wilson, R. C., & Dienst, E. R. *Evaluating university teaching.* Berkeley: Center for Research and Development in Higher Education, University of California, 1971.
Isaacson, R. L., McKeachie, W. J. & Milholland, J. E. Correlation of teacher personality variables and student ratings. *Journal of Educational Psychology*, 1963, *54*(2), 110-117.
Kulik, J. A. *Memo to the Faculty No. 58.* The University of Michigan, 1976.
Kulik, J. A. Using CRLT's instructor designed questionnaires to improve instruction. The University of Michigan, 1978. (mimeographed)
Lovell, G. D., & Haner, C. F. Forced-choice applied to college faculty rating. *Educational and Psychological Measurement*, 1955, *15*, 291-304.
Marsh, H. W., Overall, V. U., & Kesler, S. P. Class size, students' evaluations, and instructional effectiveness. *American Educational Research Journal*, 1979, *16*, 57-69.
Maslow, A. L., & Zimmerman, W. College teaching ability, activity, and personality. *Journal of Educational Psychology*, 1956, *47*, 185-189.
McKeachie, W. J., Lin, Y., & Mann, W. Student ratings of teacher effectiveness: Validity studies. *American Educational Research Journal*, 1971, *8*, 435-445.
McKeachie, W. J., & Solomon, D. Student ratings of instructors: A validity study. *Journal of Educational Research*, 1958, *51*, 379-382.
Meinkoth, M. Teachers of economic principles: Effect on student achievement and attitudes. *Journal of Experimental Education*, 1971, *40*, 66-72.
Misanchuk, E. R. *Proportionate reducation in error approach to the analysis of needs identification data.* Unpublished manuscript, University of Saskatchewan, 1981.
Morsh, J. E., Burgess, G. G., & Smith, P. N. Student achievement as a measure of instructor effectiveness. *Journal of Educational Psychology*, 1956, *47*, 79-88.
Naftulin, D. H., Ware, J. E., & Donnelly, F. A. The Dr. Fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, 1973, *48*(7), 630.
Pambookian, H. S. Initial level of student evaluation of instruction as a source of influence on instructor change after feedback. *Journal of Educational Psychology*, 1974, *66*, 52-56.
Pohlmann, J. T. A description of teaching effectiveness as measured by student ratings. *Journal of Educational Measurement*, 1975, *12*, 49-54.
Rayder, N. F. College student ratings of instructors. *Journal of Experimental Education*, 1968, *37*, 76-81.
Remmers, H. H. Appraisal of college teaching through ratings and student opinion. In *27th Yearbook of the National Society of College Teachers of Education.* Chicago: University of Chicago Press, 1939.
Rodin, M., & Rodin, B. Student evaluations of teachers. *Science*, 1972, *177*, 1164-1166.
Scissons, E. Utilizing student feedback. University of Saskatchewan Faculty/Instructional Development Seminar, March 1979.
Seibert, W. F. *National Project III: Elevating the importance of teaching.* Fund Associate's Final Report to Fund for the Improvement of Postsecondary Education, Measurement and Research Center, Purdue University, West Lafayette, Indiana, Grant No. G007502063. October, 1977.
Sheehan, D. S. On the invalidity of student ratings for administrative personnel decisions. *Journal of Higher Education*, 1975, *46*, 687-700.
Sheffield, E. F. Characteristics of effective teaching in Canadian universities—An analysis based on the testimony of a thousand graduates. *Canadian Journal of Higher Education*, 1974, STOA4, 7-30.
Shingles, R. D. Faculty ratings: Procedures for interpreting student evaluations. *American Educational Research Journal*, 1977, *14*(4), 459-470.
Starry, A. R., Derry, J. O., and Wright, G. L. An automated instructor and course appraisal system. *Educational Technology*, May, 1973, 61-64.
Stewart, C. T., and Malpass, L. F. Estimates of achievement and ratings of instructors. *Journal of Educational Research*, 1966, *59*, 347-350.
Sullivan, A. M., and Skanes, G. R. Validity of study evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology*, 1974, *66*(4), 584-590.
Treffinger, D. J., and Feldhusen, J. F. Predicting students' ratings of instruction. *Proceedings of the 78th Annual Convention of the American Psychological Association*, 1970, *5*, 621-622.
The University of Alberta. *Procedures for evaluating instruction in a university setting.* A report prepared by the Committee for the Improvement of Teaching and Learning, 1979.
The University of Massachusetts. *The clinic's teaching improvement process: some working materials.* Clinic for the Improvement of University Teaching, November, 1974. (mimeographed)
Voeks, V. W., and French, G. M. Are student-ratings of teachers affected by grades? *Journal of Higher Education*, 1960, *31*, 330-334.
Walker, B. D. An investigation of selected variables relative to the manner in which a population of junior college students evaluate their teachers. *Dissertation Abstracts*, 1969, *29*(9-B), 3474.
Wotruba, T. R., and Wright, P. L. How to develop a teacher-rating instrument: A research approach. *Journal of Higher Education*, 1975, *46*, 653-663.

*Due to the length of this list of references, the reader is encouraged to contact the author for details or copies.

The author gratefully acknowledges the contributions of Dr. Earl Misanchuk to this article, through his criticism of an earlier draft of the manuscript.