METHODS FOR IMPROVING THE INTERPRETATIVE VALUE OF STUDENT EVALUATIONS OF TEACHING

Robert D. O'Keefe, DePaul University
Lawrence 0. Hamer, DePaul University
Philip R. Kemp, DePaul University

ABSTRACT

Questionnaires that measure students' evaluations of teaching performance are widely used throughout American higher education institutions. However, while the instruments are widely used, the interpretation of the survey data is a frequent source of disagreement and discontent. This paper will discuss three methods for interpreting teacher evaluation data and suggest that interpretation is most likely to be perceived as fair and accepted when it is based upon a pre-established scale that is grounded in multiple administrations of a consistent survey instrument.

INTRODUCTION

The evaluation of teaching is one of the core activities of academic units and has profound career implications for the faculty member being evaluated. Teaching evaluation has also received increased attention as a result of at least two trends, accreditation bodies' emphasis on assessment and the Spellings Commission's emphasis on accountability (U.S. Department of Education 2006), that focusing on how both student learning and instructor effectiveness can be more effectively assessed. While few would disagree with the importance of evaluating the performance of instructors, faculty members within and across academic units tend to disagree about evaluation methods and the interpretation of the evaluation results (Clayson and Haley 1990; Kemp and O'Keefe 2003; Machina 1987; Marsh and Roche 1997; Simpson and Siguaw 2000; Theall and Franklin 1990). Fair evaluation requires a standard with which all faculty members and those charged with evaluating the performance of faculty members are familiar, find to be acceptable and consistently apply in all cases (Cranton 2001). This paper illustrates and discusses some methods which have been commonly used to compare faculty members with regard to the teaching criterion of the more comprehensive set of performance criteria traditionally considered when evaluating faculty for contract renewal, salary, and merit increases, tenure and promotion.

Higher education institutions typically use, either singularly or in combination, two types of instructor evaluation systems: peer review and student surveys. Peer reviews typically involve a faculty assessment of a particular instructor's in-class behavior, teaching materials, and/or student output. Student surveys, otherwise referred to as students evaluations of teaching or (SET) typically involve students acting as respondents to a questionnaire that assess a faculty member's performance. Regardless of the system used, the evaluation data are typically compared to some standard or expected performance in order to assess the instructor's competence. These assessments are then used to make retention, compensation, promotion, and tenure decisions. The importance of these decisions on a faculty member's career makes it imperative that the data from the evaluations are interpreted with objectivity, care and precision (Aleamoni 2000; Arreola 2000; Centra 1990; Centra and Bonesteel 2001; Ory 2000; Wergin and Swingen 2000). While there is considerable debate about SET's validity and whether or not they *should* be used (see, for a recent example, Clayson and Sheffet 2006), the fact remains that SET is used. The underlying question of this paper is that, given that SET *is* used, how can they best be interpreted? This paper will discuss various ways in which data from student evaluations of teaching are interpreted (and in some cases misinterpreted), and will suggest a method for increasing the interpretive value of teaching evaluations.

COURSE EVALUATION OBJECTIVES

SET data are typically gathered through a survey that is administered at the end of a course. The quantitative nature of the survey data eases comparison of the faculty member's performance with previous teaching performances as well as with the performance of other faculty members within and across academic units. Traditionally, course evaluation data are compared to meet either forma-

tive or summative objectives. Formative comparisons involve looking at an individual's evaluations over time to discern a trend in the evaluations (i.e., the individual's present performance is compared to his or hers past performance) (Centra 1987; Smith 2001). Summative comparisons involve looking at the individual's evaluation relative to the evaluations of his or her peers (Scriven 1987; Knapper 2000). Summative comparisons typically involve comparing a departmental average or mean score for an academic term to the individual's evaluations for that same term. Rather than the departmental mean, comparisons may be made to other faculty members of a similar rank or who teach similar material. As with all rankings and ratings, these require special attention to assure that the eventual positions ascribed and the decisions that flow from these ascribed positions are based upon transparent, objective and valid interpretation (Aleamoni 1987). Objective interpretation is especially critical when the teaching evaluation scores are submitted to the several committees and administrators who are charged with examining a faculty member's qualifications and making recommendations regarding promotion and tenure decisions.

COURSE EVALUATION COMPARISON METHODS

While evaluation comparisons are used to meet either formative or summative objectives, the actual comparisons can be made using a number of methods. Three such methods, the Observational Method, the Statistical Method, and the Empirical Method, are discussed below.

The Observational Method

The observational method of comparison uses the raw scores (means) calculated from the evaluation forms as the basis for comparison. Ease-of-use is the advantage of this technique as it requires the least amount of effort to make the comparisons. However, this technique treats scores as terminal measures and accepts any differences between scores as meaningful. Consider, for example, the evaluations presented in Table 1 for two faculty members who each taught a section of an elective undergraduate course in the same term.

The observational method would lead one to conclude that "A" is a better instructor than "B" and above average when compared to all instructors of elective undergraduate courses. Likewise, one would conclude that "B" is both worse than "A" and worse than average. Standard deviations are frequently computed and reported but, in the absence of any further statistical comparisons, the standard deviations are not much more than window dressing. The problem with the observational technique should be fairly obvious to anyone who has had a basic course in statistical analysis. Investing these raw differences with a kind of conceptual significance is the most basic of the statistical rookies' mistakes and one often used as a teaching opportunity.

The Statistical Method

The statistical method of comparison uses statistical tests to determine if differences between evaluation scores are meaningful. In this way, it overcomes the inherent error of the observational method (i.e., the assumption that all differences are meaningful). Additionally, the statistical method retains the ease-of-use advantage of the observational method as it requires only slightly more effort to compute a statistic based upon the mean and standard deviation that it does to simply observe the mean.

The primary disadvantage of the statistical technique is that it uses a relatively small amount of data in its calculations which decreases the sensitivity of the statistical tests. This makes it difficult to find statistically significant differences. Taking the data presented in Table 1, a t-test would reveal that there are no significant differences between any of the reported means. That is "A" and "B" and statistically equivalent to each other and to the departmental mean. To further illustrate this point, Table 2 presents a greater amount of data about two instructors' evaluations for a particular term along with departmental means for that term. The data in Table 2 allow each instructor to be compared to the overall departmental mean, each section to be compared to the mean for the given type of course, and for the instructors to be

TABLE 1
EXAMPLE OF OBSERVATIONAL COMPARISON METHOD

Faculty Member	N	Mean	Standard Deviation
A	29	4.5	1.1
B	32	4.2	.7
Mean of All Elective Undergraduate Courses	143	4.3	1.7

compared to each other (both overall and for a given type of course). The results of a basic t test applied to all of the possible comparisons (n = 14) indicate that the difference between each pair of means is all less than the .05 level needed to establish statistically significant differences between the scores compared. This means that, for example, an instructor scoring 3.16 on a global item has a score that does not statistically differ from the departmental mean score of 3.76. We must conclude than that, the statistical test we applied could only indicate extreme differences on a 5-point scale (i.e., the method merely distinguishes the poor teachers from the outstanding teachers). However, the method does not distinguish between differing levels of instructor performance in the broad mid range of scores.

The Empirical Method

The empirical method of comparison depends upon a longitudinal accumulation of observations derived from a consistently administered teaching evaluation instrument and the explicit agreement among those being evaluated and those charged with interpreting the ratings that the resulting outcomes are meaningful and consistently interpreted. The empirical method requires that an academic unit establish an agreed upon performance standard which its faculty members are expected to meet or exceed. Department mean scores are already used in this manner in the previously discussed observational method but these mean scores tend to change from evaluation period to evaluation period. For that reason the unit should

TABLE 2
ILLUSTRATIVE TEACHING EVALUATION DATA

	Instructor			Department		
Instructor 1	**N**	**Mean**	**Std. Dev**	**N**	**Mean**	**Std. Dev**
Overall Means						
Overall Teaching Effectiveness	61	3.69	1.07	1192	3.95	1.08
Overall Quality of the Course	59	3.75	0.9	1181	3.91	1.06
Required, Undergraduate Day						
Overall Teaching Effectiveness	29	4.28	0.59	287	4.34	0.95
Overall Quality of the Course	28	4.21	0.63	283	4.29	0.93
Elective, Graduate						
Overall Teaching Effectiveness	32	3.16	1.14	215	3.76	1.15
Overall Quality of the Course	31	3.32	0.91	212	3.69	1.09
Instructor 2						
Overall Means						
Overall Teaching Effectiveness	78	4.33	0.75	1192	3.95	1.08
Overall Quality of the Course	77	4.23	0.83	1181	3.91	1.06
Elective, Graduate						
Overall Teaching Effectiveness	32	4.67	0.68	215	3.76	1.15
Overall Quality of the Course	32	4.44	0.62	212	3.69	1.09
Elective, Undergraduate, Night						
Overall Teaching Effectiveness	46	4.11	0.82	155	3.95	1.09
Overall Quality of the Course	45	4.09	0.93	153	3.76	1.2

Note: Each of these comparisons was subjected to the traditional "t-test. We also compared the scores of Instructor 501 and Instructor 513 for an elective graduate course. Despite a number of seemingly large observational differences in comparable scores, none of the 14 possible comparisons in this table yielded statistically significant ($p < .05$) results.

choose an acceptable standard that is, in effect, based on a longitudinal "mean of means."

The advantage of the empirical technique is that it uses the greatest amount of data because it relies upon information gained from all previous evaluations rather than just the evaluations from a given term. The disadvantage of the empirical technique is that is the most difficult technique to implement. This technique requires the institution have a long history of using a particular rating scale and the perspective that comes from that scale's historical usage. Further, this technique requires that faculty agree upon the appropriate scale ranges that indicate various levels of teaching performance.

Realistically, the standard might vary with the experience level of faculty. New faculty members might be expected to meet a performance standard somewhat lower than that expected of more senior faculty. Our experience has been that such an exception is almost always implicitly granted by those charged to evaluate new faculty. A period of adjustment exception during which the comparisons are essentially formative is understandable but there must be a defined standard to be met.

The critical activity is the development of a longitudinal data base of evaluation observations of the so called "global items" for all academic periods and including all courses evaluated. While the focus remains on the "global items" the system is flexible enough to include other scale items assumed to be relevant to overall teaching performance.

We believe that an empirically-derived scale, such as that seen in Table 3, represents the resolution to problems, objections, and general expressions of discontent with the interpretation of teaching evaluation outcomes by those charged to make decisions regarding retention, compensation and the especially critical decisions bearing on recommendations for promotion and tenure.

Table 3 presents an example of an overall scale that may be used to evaluate all instructors within a given academic unit. The unit may prefer to develop multiple scales that reflect inherent differences in instructor experience or course type. For example, an academic unit observes over time that teachers of required courses are evaluated differently than those of elective courses and/or instructors of undergraduate courses are evaluated differ-

TABLE 3
EVALUATION SCALE

Unsatisfactory	Satisfactory	Good	Excellent	Outstanding
2.5 – 3.0	3.1 – 3.5	3.6 – 4.0	4.1 – 4.5	4.6 – 5.0

TABLE 4
EVALUATION SCALES THAT REFLECT DIFFERENCES IN COURSE TYPE

Required Undergraduate Courses

Unsatisfactory	Satisfactory	Good	Excellent	Outstanding
1.0 – 2.3	2.3 – 2.7	2.8 – 3.2	3.3 – 4.1	4.2 – 5.0

Elective Undergraduate Courses

Unsatisfactory	Satisfactory	Good	Excellent	Outstanding
1.0 – 2.9	3.0 – 3.4	3.5 – 3.9	4.0 – 4.4	4.5 – 5.0

Required Graduate Courses

Unsatisfactory	Satisfactory	Good	Excellent	Outstanding
1.0 – 3.0	3.1 – 3.5	3.6 – 4.0	4.1 – 4.5	4.6 – 5.0

Elective Graduate Courses

Unsatisfactory	Satisfactory	Good	Excellent	Outstanding
1.0 – 3.0	3.1 – 3.6	3.7 – 4.1	4.2 – 4.6	4.7 – 5.0

ently than those of graduate courses. Table 4 presents multiple evaluation scales that reflect these differences.

An academic unit could, in fact justify establishing a series of standards to be met for it's various programs and course levels but the more departures and exceptions incorporated into the evaluation system, the more the system departs from being a rational system and drifts into the realm of rationalization and subjectivity.

COURSE EVALUATION INTERPRETATION: SOME EXAMPLES

Interpreting course evaluations involves an interaction between a particular comparison objective and a particular comparison method. The results of this interaction are used to assess the faculty member's teaching performance and become an input into the overall assessment of teaching ability. For example, the evaluations from a particular course taught by a given faculty member may be summatively compared to evaluations from other faculty from the same academic unit using the observational method. If the given faculty member's evaluations are higher than the average evaluation from her academic unit, the given faculty member is likely to be seen as an above average instructor. However, the use of some statistical techniques may result in the misinterpretation and misassessment of a faculty member's performance. The section will discuss some appropriate and inappropriate interpretations of course evaluations and the relationship between these interpretations and the uses of evaluations.

Data Collection

Data for this study were gathered from the course evaluations administered within the Marketing department of a large, private university in the Midwest which has evaluated teaching performance with students' evaluations of courses for more than 25 years. Over time the scales have been revised from a five-point scale to a seven-point scale and then to a ten-point scale and then back again to a five-point scale, but the current scale has been in use for better than fifteen years. Scale points aside, the performance factors evaluated have remained quite consistent. The scale, which is administered to each class near the conclusion of each academic term, has seventeen bipolar scale items. There are two "global items" (see Figure 1); the first of these items asks the respondents to rate "overall teaching effectiveness and the second asks them to rate the "overall quality of the course." Eight items are directed toward the course. These items ask for ratings on matters of course organization, objectives, assignments, helpfulness of the text and whether course materials were up to date. The remaining eight items ask for ratings on questions related to the instructor. These items deal with the respondents' perceptions of whether or not the instructor was knowledgeable, explained the material well, was well organized, motivated students, fairly graded students, encouraged questions and discussion, was accessible to students, was fair and showed enthusiasm. The scale items are considered relevant to teaching performance and so the scale has, at least, face validity and the consistency of the scores achieved by faculty members over time is considered as a meaningful measure of reliability (Green, Calderon, and Powell-Reider 1995).

In addition to the scale questions the evaluation form includes on the reverse side of the scale page a series of qualitative questions dealing with perceptions of the instructor's strengths and weaknesses, the benefits of taking the course, ideas for improving the course and the fairness of the examinations and grading procedures. Each of the items is scored but the global items are given more weight. The emphasis placed on the global items is a holdover from another era of teaching evaluation instruments used within our college. In that early development period of our teaching evaluation instrument, each of the several departments within the college were allowed to include their own set of questions. The departmental representatives argued that they were better able to use questions they believed were more relevant to departmental objectives and teaching methods. All agreed to include the global items so that there would be some means for cross department comparisons. Over time the several departments agreed on the standardized form in use today. But the importance of the global items persists as a reliable and comparable summary of a faculty members teaching performance. Summaries of the results of these two items are used in comparisons with departmental means and these comparisons are included in the documentation a faculty member prepares in support of applications for promotion, tenure or both.

Appropriately Interpreting Evaluations

Because teaching evaluations are important inputs into assessment of faculty's performance, their inappropriate usage can have significant consequences on faculty satisfaction, retention, and promotion. The following are some examples of how the selection of a comparison method can lead to inappropriate comparisons.

Table 5 shows teaching evaluations received by three separate faculty members over four subsequent academic terms. The assessment of the three instructors' performances will vary depending upon the comparison method used to interpret the course evaluations.

Using the Evaluations to Meet Formative Objectives. One of the uses of teaching evaluations is to ascertain the faculty member's teaching performance over time. This formative assessment is used in the spirit of faculty development but can also be used to make compensation and retention decisions.

FIGURE 1
THE GLOBAL ITEMS AND THE SET SCALEPOINTS

Given your experience, the instructor's overall teaching effectiveness was among the worst.	1	2	3	4	5	Given your experience, the instructor's overall teaching effectiveness was among the best.
Given your experience, the overall quality of the course was among the best.	1	2	3	4	5	Given your experience, the overall quality of the course was among the best.

Using the Observational Method to Meet Formative Objectives. As the observational method treats all differences between means as significant, the use of this method to meet formative objectives merely involves observing trends in means over several terms. Thus, the data in Table 5 suggest that Instructor A's teaching as consistent improved, Instructor B's teaching has consistently deteriorated, and Instructor C's teaching is inconsistent as his performance sometimes improves and sometimes worsens. Thus, the observational method would lead to the interpretation that B is in most need of faculty development aimed at improving teaching performance while A has the least need for such development activities.

Using the Statistical Method to Meet Formative Objectives. The statistical method uses standard statistical tests to assess differences in mean scores. For the data in Table 5, the tests reveal that there are no statistically significant differences between any of the means reported in the table. In other words, this method fails to discriminate between any of the teaching performances across instructors nor does it discriminate across terms for a given instructor. Thus, interpreting the data using the statistical method would lead to the conclusion that all of the instructors are equally proficient and the teaching effectiveness of each instructor has neither improved or decreased over time.

Using the Empirical Method to Meet Formative Objectives. The empirical method involves interpreting a given teaching evaluation mean relative to a scale that is based upon historical data gathered from repeated administration of a given evaluation questionnaire and that represents the standards of evaluation agreed to by the academic unit (see Table 3). Interpreting the data in Table 5 using the standard presented in Table 3 one can see that Instructor B is consistently evaluated as "Outstanding," Instructor A is consistently evaluated as "Satisfactory," and Instructor C's evaluations vary between "Good" and "Excellent." Thus, interpreting the data using the empirical method would lead to the conclusion that Instructor A is most in need of development activities, while B is least in need of development activities.

The differing interpretations of these evaluations will lead to differing decisions about faculty development, compensation, and promotion and tenure (see Table 6).

Using Evaluations to Meet Summative Objectives

While satisfying a formative objective involves assessing an instructor's performance over time, satisfying a summative objective involves assessing an instructor's performance relative to other faculty members. This assessment is likely to be based on data from a limited point

TABLE 5
TEACHING EVALUATIONS OVER FOUR ACADEMIC TERMS

Faculty Member	Term 1	Term 2	Term 3	Term 4
A	3.18	3.26	3.30	3.41
B	4.86	4.78	4.69	4.55
C	4.02	4.11	4.06	4.09
Departmental Mean	4.11	4.03	3.97	4.15

TABLE 6
EFFECT OF EVALUATION INTERPRETATION ON FACULTY RANKINGS TO MEET FORMATIVE OBJECTIVES

	Instructor		
Comparison Method	**A**	**B**	**C**
Observation	1	3	2
Statistical		TIE	
Empirical	3	1	2

in time (e.g., a particular term or an academic year) rather than for an extended period of time. The summative evaluation objective is closely aligned with decisions regarding faculty compensation, retention, promotion, and tenure.

Using the Observational Method to Satisfy Summative Objectives. Using the observational method to make summative evaluations about faculty's teaching performance involves merely comparing the mean instructor rating of a given faculty member to the mean rating of a peer or group of peers (e.g., the mean rating for all instructors in a particular academic unit or all instructors who teach a particular course). Using the data in Table 5 for "Term 1," we can observe that both Instructors A and C are below the department mean while "B" is above the departmental mean. The interpretation of this observation is that "B" is a good instructor while "A" and "C" are not. We could also observe that C's mean is higher than that of A, and therefore conclude that C is a better instructor than A. As these evaluations are based upon comparing a faculty member to a group mean, the interpretation of the comparison assumes that the group mean is neutral (i.e., the mean does not represent "good" performance or poor performance). This could result in a particular faculty member's teaching performance remaining constant but being differently evaluated as a result of variations in the group mean from one evaluation period to another.

Using the Statistical Method to Meet Summative Objectives. The statistical method is likely to be unsuited for making summative evaluations as the method only detects extreme difference in teaching performances. As mentioned earlier, all of the means presented in Table 2 are statistically equivalent, so no inferences about the relative quality of the instructors can be made using statistical tests (i.e., each of the instructors would be interpreted as being equivalent to the other instructors).

Using the Empirical Method to Meet Summative Objectives. The empirical method involves interpreting teaching evaluation means for the various instructors to a predetermined scale (e.g., Table 3). Again using the "Term 1" data from Table 5, this comparison reveals that instructor B is "Outstanding," C is "Good," and A is "Satisfactory." Compensation decisions based upon this data could reflect the interpretation that B is at the top of the teaching scale, while A has a great deal of room for improvement. Further, these interpretations are more stable as the comparison standard is stable over multiple time periods rather than changing with each term. Thus, the interpretations are likely to be seen by the faculty as more fair and indicative of their true teaching performance.

Looking across the three different comparisons methods, it can be seen that the choice of comparison method effects the interpretation of the instructors' evaluations when using the evaluations to meet summative objectives (see Table 7).

It is interesting to note that while the observation and empirical methods result in the same ranking of instructors, they do not result in the same information about the instructors' effectiveness. While, the observational method ranks the instructors, thus establishing the relative abilities of the instructors, it does not give information about the objective abilities. The individual interpreting the course evaluations would have no way of knowing if any or all of the instructors is excellent, poor, or somewhere between these two extremes. However, the empirical method does provide information about each instructor's objective abilities and this information is used to form the rankings presented in Table 7.

Alternative Comparisons

Faculty members are frequently concerned with the comparison standard used in the course of performance evaluations. The tendency on the part of some faculty members and academic units is to change the unit of comparison in a way that is perceived to be more fair to the faculty member (i.e., choose a unit of comparison that better represents the circumstances of the faculty mem-

TABLE 7
EFFECT OF EVALUATION INTERPRETATION ON FACULTY RANKINGS TO MEET SUMMATIVE OBJECTIVES

	Instructor		
Comparison Method	**A**	**B**	**C**
Observation	Poor	Good	Poor
Statistical		Equal	
Empirical	Satisfactory	Outstanding	Good

ber). Thus, depending upon the faculty member and course of interest, the faculty member's performance can be compared to:

1. All instructors in the academic unit.
2. All instructors who teach the same course.
3. All instructors who teach the same type of course (elective or required, undergraduate or graduate).
4. All instructors of the same rank or level of experience.
5. All instructors with the same course load.

While these alternative comparison populations may provide more information for a faculty member to use when assessing her own performance, they also further strengthen the need for an agreed upon evaluation scale such as that provided in Table 3. The existence of multiple comparison populations increases the chance that a faculty member's performance will be interpreted differently by the many committee members, board members, administrators, etc. who are charged with judge teaching performance. This is especially problematic in promotion and tenure decisions as such decisions involve a greater number of judges who typically come from a greater variety of backgrounds. Reporting the faculty member's teaching performance as defined by an agreed upon evaluation scale will result in true increases in the "fairness" of the promotion and tenure process.

SUMMARY

This article discussed a number of issues relevant to the interpretation of the results derived from student teaching evaluations. It presented three models: the Observational, the Statistical, and the Empirical which are applied when making comparisons between the performance of individual faculty members or between the performance of a single faculty member compared with all other members of an academic unit within a given time frame. The article includes examples of the interpretations which result from the application of each of the three methods in rating faculty members' teaching performance.

The empirical model for comparisons which we suggest academic units adopt as an evaluative format can be applied to both the summative and formative approaches. Canon (2001) wrote that the interest in teaching evaluation was engendered by the evaluation practices commonly used in business. Interestingly enough while academics have focused largely on the summative approach businesses have recently turned more toward the formative approach. Writing in *Fortune,* Betsy Morris (2006) explained the General Electric under the leadership of Jack Welsch made summative evaluations of his managers. Under the new leadership CEO Jeffrey Immeldt has abandoned the summative method in favor of the formative method in which a manager's performance in a given period is compared to his or her performance over time. Also a recent study (McGregor 2006) reports that the multi item and frequently lengthy questionnaires requesting customers to evaluate services could be replaced by a single question survey. This question suggested as a "Global Item" asks, that on a scale of 1 to 10, people report: how strongly they would recommend the service to other people. Perhaps, this sort of scale could be used as a substitute or validation when collecting data concerning the so called global items regarding the effectiveness of the instructor and the quality of the course.

REFERENCES

Aleamoni, L.M. (1987), "Typical Faculty Concerns About Student Evaluation of Teaching," in *Techniques for Evaluating and Improving Instruction,* L.A. Aleamoni, ed. New Directions for Teaching and Learning, #31. San Francisco: Jossey-Bass.

Arreola, R.A. (2000), *Developing a Comprehensive Fac-*

ulty Evaluation System. Bolton, MA: Anker Publishing Co.

Canon, R. (2001), "Broadening the Context for Teaching Evaluation," in *Fresh Approaches to the Evaluation of Teaching. New Directions for Reaching and Learning,* C. Knapper and P. Cranton, eds #88, San Francisco: Jossey-Bass.

Centra, J.A. (1990), "Evaluating College Teaching: Some Reflections," Departmental Advisor, 5 (Winter), 1–5

_____ (1997), "Formative and Summative Evaluation: Parody or Paradox?" in *Techniques for Evaluating and Improving Instruction,* L.M. Aleamoni, ed. New Directions for Teaching and Learning, #31, San Francisco: Jossey-Bass.

_____ and P. Bonesteel (2001), "College Teaching: An Art or a Science," in *Student Ratings of Instruction: Issues for Improving Practice. New Directions for Teaching and Learning,* M.Theall and J. Franklin, eds. #43, San Francisco: Jossey-Bass.

Clayson, Dennis E. and Mary Jane Sheffet (2006), "Personality and the Student Evaluation of Teaching," *Journal of Marketing Education,* 28 (2), 149–60.

Cranton, P. (2001), "Interpretive and Critical Evaluation," in *Fresh Approaches to the Evaluation of Teaching. New Directions for Teaching and Learning,* C. Knapper and P. Cranton, eds. #88, San Francisco: Jossey-Bass.

Green, B.P., T.G. Calderon, and B. Powell-Reider (1995), "A Content Analysis of the Validity of Teaching Evaluation Instruments Used in Accounting Departments. Presentation," *American Accounting Association's National Meeting.*

Hayes, E. (1989), "Editors Notes," in *New Directions for Continuing Education: Effective Teaching Styles. New Directions for Teaching and Learning,* E. Hayes, ed. #43 San Francisco: Jossey-Bass.

Kemp, P and R.D. O'Keefe (2003), "Improving Teaching Effectiveness: Some Examples from a Program for the Enhancement of Teaching," *College Teaching,* 51 (3), (Summer), 111–14.

Knapper, C. (2001), "Broadening Our Approach to Teaching Evaluation," in *Fresh Approaches to the Evaluation of Teaching. New Directions for Teaching and Learning,* C. Knapper and P. Cranton, eds. #88. San Francisco: Jossey-Bass.

McGregor, J. (2006), "Would You Recommend Us?" *Business Week,* (January 30), 94–95.

Morris, Betsy (2006), "The New Rules," *Fortune,* (July 24), 70–87.

Ory, J.C. (2000), "Teaching Evaluation: Past, Present and Future," in *Evaluating Teaching in Higher Education: A Vision for the Future. New Directions for Teaching and Learning,* K. Ryan, ed. #83. San Francisco: Jossey-Bass.

Scriven, M. (1981), "Summative Teacher Evaluation," in *Handbook of Teacher Evaluation,* J. Millman, ed. Beverly Hills: Sage Publishing.

Smith, R. (2001), "Formative Evaluation and the Scholarship of Learning," in *Fresh Approaches to the Evaluation of Teaching. New Directions for Teaching and Learning,* C. Knapper and P. Cranton, eds. #88, San Francisco: Jossey Bass.

Symonds, W. and R. Miller (2002), "Harvard," (Cover Story), *Business Week,* (February 18), 72–78.

Theall, M. and J. Franklin (1990), "Student Ratings in the Context of Complex Evaluation Systems," in *Student Ratings of Instruction: Issues for Improving Practice. New Directions for Teaching and Learning,* M. Theall and J. Franklin, eds. #43, San Francisco: Jossey-Bass.

_____ and _____ (2000), "Creating Responsive Student Ratings Systems to Improve Evaluation Practice," in *Evaluating Teaching in Higher Education: A Vision for the Future. New Directions for Teaching and Learning,* K. Ryan, ed. #83, San Francisco: Jossey-Bass.

Underwood, B., C. Duncan, A. Taylor, and J. Cotton (1954), *Elementary Statistics.* New York: Appleton-Century Crofts Inc.

U.S. Department of Education (2006), *A Test of Leadership: Charting the Future of U.S. Higher Education.* Washington D.C.

Wergin, J. and J. Swingen (2000), *Departmental Assessment: How Some Campuses are Efficiently Evaluating the Collective Work of Faculty.* Washington D.C.: American Association for Higher Education.