Assessment & Evaluation in Higher Education
Vol. 33, No. 5, October 2008, 567–576

Routledge
Taylor & Francis Group

Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness

Charles A. Burdsal* and Paul D. Harrison

Department of Psychology, Wichita State University, Wichita, KS, USA

The purpose of this research is to provide additional empirical evidence supporting the use of both a multidimensional profile and an overall evaluation of teaching effectiveness as valid indicators of student perceptions of effective classroom instruction. A factor analytic teaching evaluation instrument was used that also included open-ended comments on four questions. Numerical scores from 208 classes were matched with the average valence of the open-ended comments. It was found that the average valences were highly positively correlated with the numerical factor scores that make up the multidimensional profile of teaching effectiveness and with the second-order factor that serves as an overall evaluation of teaching effectiveness. The implications of these results for the usefulness of student evaluations are discussed.

Introduction

Prior research has debated the merits of using an overall evaluation of teaching effectiveness versus a multidimensional profile of teaching effectiveness for summative purposes (i.e. personnel decisions). Abrami and his colleagues (Abrami 1985, 1989; Abrami and d'Apollonia 1999; d'Apollonia and Abrami 1997) feel a single overall assessment of teaching should be employed by averaging responses to several global items. They further argue that there is a practical advantage to this approach because summative decisions are almost always unidimensional. Frey (1973, 1974, 1978) argues strongly that only individual teaching dimensions should be considered, and he excluded global rating items from his Endeavor instrument. Marsh and his colleagues have chosen a middle ground between the positions of Abrami and Frey, recommending the use of both individual teaching dimensions and global ratings (Marsh 1987, 1989, 1991a; Marsh and Dunkin 1992). Consistent with the position taken by Marsh and his colleagues, Ryan and Harrison (1995) recommended that three types of student rating information be used in making personnel decisions: (1) individual teaching dimension ratings, (2) overall evaluations made by students, and (3) a composite weighted average overall evaluation.

The question that naturally arises is whether or not the students' real perceptions concerning the teaching performance are accurately reflected by the measures in the student evaluation instrument being used. The objective of this research is to provide empirical evidence as to the efficacy of using a multidimensional profile and an overall evaluation of teaching effectiveness as valid indicators of student perceptions of teaching effectiveness. We will accomplish this objective by providing empirical evidence which shows that the valence of the open-ended comments on a reliable factor-analytic student evaluation of teaching effectiveness instrument (SETE) received by individual faculty members is highly correlated with both the individual teaching dimensions on the SETE (the first-order factors) and the overall evaluation provided by the SETE.

*Corresponding author. Email: charles.burdsal@wichita.edu

ISSN 0260-2938 print/ISSN 1469-297X online

DOI: 10.1080/02602930701699049
http://www.informaworld.com

Open-ended comments

Braskamp et al. (1980) looked at whether the summary judgments of teaching performance and course quality based on student written comments and group interviews are convergent with the overall evaluations on a student evaluation instrument. The Pearson product moment correlations between all pairs of average course and instructor ratings was 0.86, indicating a high degree of similarity between these three different measures of overall quality.

Ory and Braskamp (1981) investigated faculty perceptions of three different types of student evaluation information: objective questionnaire items, open-ended student comments, and group interviews. In general, faculty regarded the information to be more credible, useful and accurate for their own self-improvement than for promotion purposes. They regarded student written comments as less credible than student responses to objective questions when the purpose was promotion, but rated written comments as more credible when the purpose was self-improvement. Furthermore, faculty desired more than one type of evaluative information regardless of the purpose of evaluation.

Overall vs. multidimensional evaluations

A major issue frequently debated in teacher evaluation research and practice deals with the relative merits of using an overall evaluation versus a multidimensional profile of teaching effectiveness. For personnel decisions, there is considerable debate as to whether a single score is more useful and appropriate than a profile of scores reflecting multiple dimensions (see Abrami 1989; Abrami and d'Apollonia 1991; Cashin and Downey 1992; Marsh 1987, 1989, 1991a; Marsh and Hocevar 1991).

Abrami and his colleagues (Abrami 1985, 1989; Abrami and d'Apollonia 1990) favour the use of several global items to evaluate teaching for personnel decisions. They have argued against the use of separate factor scores for personnel decisions. First, they are not convinced that any of the carefully developed, well-validated rating forms represent the teaching dimensions invariantly. They failed to find evidence of the replicability of teaching factors across rating forms (Abrami and d'Apollonia 1990). Second, they are concerned about the content validity of specific items and some of the dimensions they compromise when ratings are used across a wide variety of courses, instructors, students and settings. Third, Abrami and his colleagues feel that Cohen's (1981) review of multi-section validity studies suggests that many rating dimensions have lower correlations with student learning than with overall course and overall instructor ratings. Fourth, less is known about the generalisability of specific factors than overall ratings. Finally, researchers have concerns about the ability of administrators or non-experts to properly weigh the information provided by factor scores in arriving at a single decision on the quality of good teaching (Franklin and Theall 1989).

Conversely, Frey (1973, 1974, 1978) argues strongly that only individual teaching dimensions should be considered, and he excluded global rating items from his Endeavor instrument. His subsequent research on two higher-order dimensions (Frey and Flay 1978) led him to conclude 'that personnel decisions should not be made on a single global evaluation measure' (Frey and Flay 1978, 25). Frey's main arguments were that: (a) global items are too much influenced by variables that are not associated with effective teaching, (b) global ratings are unduly influenced by student evaluation of teaching effectiveness (SETE) components that are minimally related to student achievement, and (c) it is better to focus on components that are maximally related to a particular criterion than to rely on global items (Marsh 1991b).

Marsh and his colleagues (Marsh 1987, 1989, 1991a; Marsh and Hocevar 1991; Marsh and Dunkin 1992; Marsh and Roche 1997, 1999) have argued that teaching is multidimensional and that the individual teaching dimensions should be considered in evaluating teaching effectiveness.

They have chosen a middle ground between the positions postulated above, recommending the use of both specific dimensions and global ratings in personnel decisions. They believe that the overall rating should be a weighted average of the individual factors, with the weights being determined by logical and empirical analysis (Marsh 1991b, 1994; Marsh and Hocevar 1991; Marsh and Roche 1992, 1997). These weights could be constructed on the basis of empirical research findings or ratings of the relative importance of specific components by the department head, a promotions committee, or the instructor (Marsh 1991b). Marsh further notes that the use of weighted averages is 'a compromise that seems consistent with recommendations by Abrami, Frey and myself' (Marsh 1991b, 419).

Weighted average overall evaluations

Feldman's (1988) review of prior research shows that the correlation between students' specific evaluations of teachers on individual instructional dimensions and overall evaluations of teachers reflects the relative importance of various specific instructional characteristics in discriminating among students' global assessment of teachers. Feldman indicates that:

> The teacher's preparation and organization, clarity and understandableness, and sensitivity to, and concern with, class level and progress are of especial importance in all three ways. That is, students and faculty view them as highly important when asked about the components of good teaching, and they are of high importance in discriminating among the global ratings received by teachers from their students. (1988, 316)

However, this analysis does not provide a direct measure of the relative importance of different factors in arriving at an overall evaluation of teaching effectiveness.

Marsh and Roche (1993) evaluated the effectiveness of students' evaluations of teaching effectiveness (SETEs) as a means for enhancing university teaching. The teacher being evaluated weighted the individual dimensions in Marsh's SEEQ with regard to relative importance. Consistent with their expectation, the weighted average overall evaluation improved when there was individually structured intervention targeted at specific dimensions of the ASEEQ (Australia version).

Ryan and Harrison (1995) provide experimental evidence of how students implicitly weight various teaching factors in arriving at an overall evaluation of teaching effectiveness. A policy-capturing experiment was conducted in which students in three different instructional contexts (accounting, education and geology) made overall evaluations of hypothetical instructors based on an orthogonal manipulation of the teaching factors in Marsh's SEEQ (Marsh 1982, 1983). The teaching factors in the SEEQ are (1) Learning, (2) Enthusiasm, (3) Organisation, (4) Group Interaction, (5) Individual Rapport, (6) Breadth of Coverage, (7) Examination Fairness, (8) Assignments, and (9) Course Difficulty. The results indicated: (1) Amount Learned was consistently the most important factor affecting overall evaluations; (2) Course Difficulty was consistently the least important factor affecting overall evaluations; and (3) there was a strong similarity among the three groups in the relative importance of the various teaching factors in arriving at an overall evaluation.

In the same experimental study, Harrison et al. (1996) asked these three groups of students how important each of the teaching factors in Marsh's SEEQ should be in making an overall evaluation of teaching effectiveness. These explicit weights were then correlated with the implicit weights reported in Ryan and Harrison (1995) and it was determined that these three groups of students: (1) relied upon a common implicit theory, (2) had a reasonably high level of consensus, and (3) demonstrated self-insight when making their overall evaluations of the hypothetical instructors. These two studies collectively indicate that there is a direct link between how students made overall evaluations of hypothetical instructors and their opinions on how important

individual teaching factors should be when making an overall evaluation. This demonstrated link would become important if a weighted average overall evaluation based on student-derived weights were to be used.

Harrison et al. (2004) compared the relative usefulness of different types of overall evaluations of teaching effectiveness in a classroom setting. Using a norm-based factor analytic instrument (*Student Perceptions of Teaching Effectiveness, SPTE*), Harrison et al. (2004) compared the merits of five different types of overall evaluations: (1) a student-derived weighted average overall evaluation, (2) a faculty-derived weighted average overall evaluation, (3) an unweighted average overall evaluation, (4) an overall evaluation made by students, and (5) the second-order factor on the *SPTE* which proxies for an overall evaluation, the *Perceived Quality Index* (*PQI*). Their results indicated that: (1) all of these overall evaluations of teaching effectiveness were highly intercorrelated (beyond 0.81), and (2) the second-order factor (*PQI*) that serves as an overall evaluation on the *SPTE* was the most highly correlated with the other overall evaluations of teaching effectiveness and had the further advantage of being most understandable by the faculty. The results further indicated that: (1) the four first-order factors that load on the *PQI* were also highly correlated with the overall evaluation made by the students, and (2) the two first-order factors which do not load on the *PQI*, but instead load on the second-order factor *Course Demands* had very low correlations with the overall evaluation made by the students.

The purpose of this research is to provide further evidence that both a multidimensional profile and an overall evaluation of teaching effectiveness are valid indicators of student perceptions of teaching effectiveness. We will accomplish this by comparing the valence of the student open-ended comments on a factor-analytic student evaluation of teaching effectiveness instrument with both the first-order factors of teaching effectiveness (the multidimensional profile), and the overall evaluation (the second-order factor that serves as an overall evaluation on this particular SETE) of teaching effectiveness.

Method

Student evaluation instrument

The *Student Perceptions of Teaching Effectiveness* (*SPTE*) is a norm-based instrument. Using factor-analytic statistical techniques, six first-order factors and two second-order factors have been identified. Four of the six first-order factors are: *Course Organisation and Design*, *Rapport with Students*, *Grading Quality* and *Course Value*. These four first-order factors comprise the multidimensional profile of teaching effectiveness for the instructor. They load on the second-order factor *Perceived Quality Index* (*PQI*), which is used as an overall evaluation with this particular instrument (see the literature review above). The other two first-order factors are *Course Difficulty* and *Workload*, and these two factors load on the second-order factor *Perceived Course Demands*. With a database of more than 8000 classes, these factors have been found to be both reliable and stable (Burdsal and Bardo 1986; Jackson et al. 1999).

Summative and formative information is returned to the instructor after the semester is over. The front page contains normed results for each of the eight factor scales discussed above. The top box contains the instructor's scores on the *Perceived Quality Index* (*PQI*), as well as the scores on the four first-order factors that make up the *PQI*: *Course Organisation and Design*, *Rapport with Students*, *Grading Quality* and *Course Value*. The lower box includes the instructor's scores on the second-order factor *Perceived Course Demands*, and the two first-order factor scores that make up this scale, *Course Difficulty* and *Workload*. The *PQI* and *Perceived Course Demands* are virtually orthogonal ($r = 0.04$). Two numbers are given for each scale, a *SCALE* score (the *SCALE* score is adjusted to have a mean of 5.5 and a standard deviation of 2), and the

PERCENTILE. Additional information consisting of item scores, both raw and standardised, is found on the back. All scores are corrected for a priori motivation and class size (Marsh 1982; Marsh and Roche 1997).

In addition, the students filled out a *SPTE* comment sheet. The four questions that students are free to respond to are: (1) What could the instructor do to improve the course? (2) What did you like about the course/and or instructor? (3) Please comment on the effectiveness of computer-aided instruction. (4) Add any additional comments.

Sample

At the beginning of the data collection, 1086 classes requested the *SPTE* to be administered. From the list of the requests, every other instructor was selected to participate. If the instructor had requested the *SPTE* for more than one class, one class was randomly selected. This resulted in an initial sample of 250 classes. Of these classes, 25 chose to not participate, and 13 were excluded because of minor problems (i.e. cancellation of the evaluation, member of the research group, etc). This left us with a sample size of 212 classes.

The Associate Director of the Social Science Research Lab (which administers the *SPTE*) copied all of the comments and returned the originals to the individual faculty members. She then read all of the comments, eliminating anything (e.g. darkening out the instructor's name) that might identify the faculty member. The comment sheets were then numbered and packaged in class-identifying packets. There were a total of 3274 comment sheets.

Demographic information concerning the student sample was collected from the *SPTE* forms. The average class composition was 45.9% male and 54.1 % female. Of the students in the classes 12.7 % were freshmen, 16.2 % sophomores, 23.1 % juniors, 25.8% seniors and 22.3% were graduate students.

Data collection and initial analysis

As the responses students gave to each item on the open-ended questionnaire often contained more than one thought, an initial review of all the questionnaires was done to divide the answers into individual thoughts. The comments were divided if there was more than one complete thought. The comments were then typed into an Excel spreadsheet.

After this procedure was completed for all of the classes, team members reviewed comments they had not entered. The total number of unitised comments (across all classes) was 14,312.

Rating comment valences

The next step was to rate the favourableness–unfavourableness of the comments. After examining a sample of the comments, a five-point classification scheme was developed with 1 being the most negative, 3 being neutral and 5 being the most positive.

Twelve raters were then divided into two teams, A and B. One person from team A was matched with a person from Team B. Each pair rated the same class's (in their own class-identifying packets) comments for valence. Cohen's Kappa (Cohen 1960) was computed to establish the inter-rater reliability between the two raters. Cohen's Kappa for the valences was computed to be 0.823, which is quite acceptable for this type of research (Cohen 1960).

A list was generated for disagreements on the assigned valences. All differences were put in a separate Excel file. Three teams of three each were assigned one-third of the list. The teams discussed the differences, and these differences were resolved by consensus. One Excel file was then created with the final valences for all the comments.

Results

The average valence (favourableness, unfavourableness) scores were expected to vary by question on the *SPTE* comment sheet. The first question asked: 'What could the instructor do to improve the course?' The second question asked: 'What did you like about the course/and or the instructor?' By the very nature of these two questions, we expected more unfavourable comments on the first question and more favourable comments on the second question. The third question was: 'Please comment on the effectiveness of computer-aided instruction.' The fourth question was: 'Add any additional comments.' We expected the average valence to be more in the mid-range on these two questions, due to their more neutral phrasing.

Table 1 gives the distribution of responses: (1) across the five possible valences, (2) across all four questions, and (3) across the five possible valences for each question As Table 1 indicates, the majority of the comments were positive, with 60% of the comments in the positive range, 34% of the comments in the negative range, and 6% neutral. The mean valence was 3.4, with a standard deviation of 0.377. The mean valence for each class ranged from 2.33 to 4.17. As predicted, a majority of the comments for question one were in the negative range (78%), and a majority of the comments were in the positive range (94%) for question 2; 58% of the comments were in the positive range for question 3, and 51% of the comments were in the positive range for question 4.

In analysing the relationship between the open-ended comment valences and the *SPTE* factor scales, we first computed a mean valence for the comments by question for each section, and an overall mean valence for all the comments for a section. We then correlated the mean valence for each question with each of the six first-order factors and the two second-order factors on the *SPTE*. We also computed the correlation between the mean overall valence for each section and each of the six first-order factors and the two second-order factors on the *SPTE*. Four classes had too few students for scale scores to be calculated, resulting in 208 classes for which we had complete data.

Table 2 gives the correlations described above. First, the mean valence per section for questions 1, 2 and 4 (see Table 1 for the questions) have high positive correlations with the four first-order factors that serve as the multidimensional profile of teaching effectiveness with this particular *SETE* instrument. These correlations are all at 0.508 or higher. Second, the mean valences per section for each of these three questions are all highly positively correlated with the second-order factor that serves as an overall evaluation with this particular instrument, the *Perceived Quality Index* (*PQI*) (0.603 or higher). Third, the overall mean valence per section is

Table 1. Frequency of valences by question number.

	Question number				
Valence	(1) What could the instructor do to improve the course?	(2) What do you like about the course/ instructor?	(3) Please comment on the effectiveness of computer-aided instruction.	(4) Add any additional comments	Total
1 (neg/neg)	37 (.3%)	23 (.2%)	11 (.1%)	67 (.5%)	138 (1.0%)
2 (negative)	3,331 (23.3%)	223 (1.6%)	490 (3.4%)	681 (4.8%)	4,725 (33.0%)
3 (neutral)	121 (.8%)	69 (.5%)	511 (3.6%)	202 (1.4%)	903 (6.3%)
4 (positive)	695 (4.9%)	4,934 (34.5%)	1,250 (8.7%	726 (5.1%)	7,605 (53.1%
5 (pos/pos)	158 (1.1%)	370 (2.6%)	158 (1.1%)	255 (1.8%)	941 (6.6%
Total	4,342 (30.3%)	5,619 (39.3%)	2,420 (16.9%)	1,931 (13.5%)	14,312 (100%)

Table 2. Mean valences (n = 196–208).[1]

Scale	(1) What could the instructor do to improve the course?	(2) What do you like about the course/ instructor?	(3) Please comment on the effectiveness of computer-aided instruction	(4) Add any additional comments.	Overall	Overall without Item 3
Rapport	0.574	0.548	0.225	0.634	0.736	0.750
Course value	0.609	0.572	0.277	0.638	0.768	0.778
Course design	0.545	0.599	0.252	0.599	0.737	0.740
Grading quality	0.581	0.508	0.324	0.634	0.738	0.721
Perceived quality Index	0.603	0.603	0.286	0.664	0.790	0.792
Difficulty	–0.283	0.008	–0.078	–0.169	–0.205	–0.210
Workload	–0.268	–0.223	–0.080	–0.221	–0.276	–0.295
Course demands	–0.293	–0.058	–0.078	–0.194	–0.234	–0.244

Note: [1]There is a slight variation in n depending on the question and section. For some sections there were no comments for one or more of the questions.

highly positively correlated with both the four first-order factors that serve as the multidimensional profile of teaching effectiveness (0.736 or higher) and with the *PQI* (0.790).

Several additional points are noteworthy. First, the mean valence per section for question 3 has much lower correlations with both the four first-order factors and the second-order factor (the *PQI*) when compared with the three other open-ended questions. Second, the overall mean valence per section for each of the four open-ended questions has, for the most part, relatively low negative correlations with the two other first-order factors (*Course Difficulty* and *Workload*) and the second-order factor on which they load (*Course Demands*). Third, when question 3 is omitted from the analysis, the overall mean valence per section has higher positive correlations with three of the four first-order factors that make up the multidimensional profile of teaching effectiveness (the exception being *Grading Quality*) and with the overall evaluation (*PQI*). This is to be expected considering the relatively low positive correlations question 3 has with these scales. (See Table 2 for both the overall correlations without item 3 and the correlations of item 3 with the *SPTE* scales.)

Discussion and conclusions

The objective of this research was to provide empirical evidence as to the efficacy of using a multidimensional profile and an overall evaluation of teaching effectiveness as reliable indicators of student perceptions of teaching effectiveness. Using the *Student Perceptions of Teaching Effectiveness* (*SPTE*) factor-analytic student evaluation of teaching effectiveness (SETE) instrument, we found that the mean valence per section for three of the four open-ended questions had a strong positive correlation both with the four first-order factors that serve as a multidimensional profile of teaching effectiveness, and with the second-order factor (the PQI) that serves as an overall evaluation with this SETE instrument. We also found that the overall mean valence per section of the written comments had a strong positive correlation with both the multidimensional profile of teaching effectiveness and the overall evaluation used with this SETE. These results collectively indicate that the higher the average valence on the open-ended questions in a given section, the higher the student ratings tended to be on both the multidimensional profile (the four first-order factors) and the second-order factor that serves as an overall evaluation of teaching

effectiveness (the *PQI*). There were also mildly negative correlations between the average valence per section of the open-ended questions and the two other first-order factors, Difficulty and Workload, and the second-order factor they load on, Course Demands. This relationship also existed with the overall mean valence per section and the two first-order factors, Difficulty and Workload, and the second-order factor, Course Demands. This result indicates that the more positive the average valence of the open-ended comments are in a section, the less difficult and demanding a class is perceived to be. These effects, however, were small accounting for at most 8.6% to less than 1% of the variance. These results held when we computed the mean valence for the comments by question and overall mean valence for each section.

Question 3 on the *SPTE* comment sheet was 'Please comment on the effectiveness of computer-aided instruction.' The correlations of the mean valence per section on this question with both the four first-order factors that serve as a multidimensional profile of teaching effectiveness and the second-order factor that serves as an overall evaluation with this particular SETE was much lower than that of the other three open-ended questions. These results seem to imply that the comments on this question are related to the use of computer technology, and that the students appear to view this separately from classroom teaching performance. Further research is needed to investigate why this is the case. For the vast majority of classes at this university, the use of computer technology is mainly related to using Blackboard to post grades in a timely manner, and to make class handouts, outlines, etc. available to students online. As the use of computer technology is incorporated more thoroughly into the curriculum (i.e. PowerPoint presentations, internet-based research, etc.) the correlation of the valence of this open-ended question with the multidimensional profile of teaching effectiveness and with the overall evaluation may very well increase.

Future research is needed to determine whether these results generalise to other reliable factor-analytic student evaluation of teaching effectiveness instruments. Specifically, if another factor-analytic instrument has a second-order factor that proxies for an overall evaluation as the *PQI* does on the *SPTE*, it would be worthwhile to know if this second-order factor (and the first-order factors that load on it) are highly correlated with student attitudes (via the student comments) concerning the effectiveness of the university instruction. Alternatively, if a student evaluation instrument has the students make an overall evaluation, the future research could investigate the relationship between this type of overall evaluation and student attitudes to instruction. The work of Harrison et al. (2004) would suggest that there would be a strong association. The correlation between the *PQI* and the overall evaluation made by the students in their research was 0.903.

We believe that SETE instruments measure students' *attitudes* to teaching effectiveness, not necessarily teaching effectiveness *per se*. That being said, we believe the results of this study provide evidence of concurrent validity for this particular SETE. That is, using the *SPTE*, we have provided additional empirical evidence that the multidimensional profile of teaching effectiveness and the overall evaluation of teaching effectiveness used here are valid measures of student perceptions of teaching effectiveness in the classroom.

A reliable and valid measure of a particular group's perceptions regarding effective teaching is not the same thing as having a valid measure of effective teaching. Given the wide variety of courses taught in a university, we are not aware of any definition of teaching that is acceptable and valid across all disciplines in a university setting. However, we also know that at this particular university students will not sign up for courses taught by instructors with the worst multidimensional profiles of teaching effectiveness and with the lowest overall evaluations (the *PQI*) unless they are somehow forced into these classes (i.e. required class for a major, and nobody else teaches it). For this reason, we join others who recommend that a teaching portfolio be developed for university professors (Seldin 1991). This portfolio should

have many indicators of their teaching performance. Student evaluations should be just one component of this portfolio.

Notes on contributors

Charles A. Burdsal is professor and chair of psychology and director of the Social Science Research Laboratory. His current research interests are teaching evaluation, programme evaluation and multivariate methodology.

Paul D. Harrison is professor of accountancy. His current research interests are teaching evaluation, escalation of commitment, and financial statement fraud.

References

Abrami, P.C. 1985. Dimensions of effective college instruction. *Review of Higher Education* 8, no. 3: 211–28.

———. 1989. How should we use student ratings to evaluate teaching? *Research in Higher Education* 30, no. 2: 221–7.

Abrami, P.C., and S. d'Apollonia. 1990. The dimensionality of ratings and their use in personnel decisions. In *New directions for teaching and learning student ratings of instruction: issues for improving practice,* no. 43, eds. M. Theall and J. Franklin, 97–111. San Francisco: Jossey-Bass.

———. 1991. Multidimensional students' evaluations of teaching effectiveness–generalizability of 'N = 1' research: Comment on Marsh (1991). *Journal of Educational Psychology* 83, no. 3: 411–5.

———. 1999. Current concerns are past concerns. *American Psychologist* 54, no. 7: 519–20.

Braskamp, L.A., J.C., Ory, and D.M. Pieper. 1980. Congruency of student evaluative information collected by three methods. *Journal of Educational Psychology* 72, no. 2: 181–5.

Burdsal, C.A., and J.W. Bardo. 1986. Measuring students' perceptions of teaching: dimensions of evaluation. *Educational and Psychological Measurement* 46: 63–79.

D'Apollonia, S., and P.C. Abrami. 1997. Navigating student ratings of instruction. *Journal of Educational Psychology* 52, no. 11: 1198–208.

Cashin, W.E., and R.G. Downey. 1992. Using global student rating items for summative evaluation. *Journal of Educational Psychology* 84, no. 4: 563–72.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, no. 1: 37–46.

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.

Feldman, K.A. 1988. Effective college teaching from the students' and faculty view: matched or mismatched priorities? *Research in Higher Education* 28, no. 4: 291–344.

Franklin, U., and M. Theall. 1989. Rating the readers: knowledge, attitude, and practice of users of student ratings of instruction. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Frey, P.W. 1973. Student ratings of teaching: validity of several rating factors. *Science* 182: 83–5.

———. 1974. The ongoing debate: student evaluation of teaching. *Change* February: 47–49.

———. 1978. A two-dimensional analysis of student ratings of instruction. *Research in Higher Education* 9, no. 1: 69–91.

Frey, P.W. and Flay, B.R. 1978. *A cusp catastrophe model of evaluation person perception with an application to student ratings of instruction.* Evanston, IL: Northwestern University.

Harrison, P.D., J.M. Ryan, and P.S. Moore. 1996. College students' self-insight and common implicit theories in ratings of teaching effectiveness. *Journal of Educational Psychology* 88, no. 4: 775–82.

Harrison, P.D., D.K. Douglas, and C.A. Burdsal. 2004. The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education* 45, no. 3: 311–23.

Jackson, D.L., C.R. Teal, S.J. Raines, T.R. Nansel, R.C. Force, and C.A. Burdsal. 1999. The dimensions of students' perceptions of teaching effectiveness. *Educational and Psychological Measurement* 59, no. 4: 580–96.

Marsh, H.W. 1982. SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology* 52, no. 1: 77–95.

———. 1983. Multidimensional ratings of teaching effectiveness by students for different academic settings and their relationship to student/course/instructor characteristics. *Journal of Educational Psychology* 75: 150–66.

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

———. 1989. Responses to reviews of students' evaluations of university teaching: research findings, methodological issues, and directions for future research. *Instructional Evaluation* 10: 5–9.

———. 1991a. Multidimensional students' evaluations of teaching effectiveness: a test of higher-order structures. *Journal of Educational Psychology* 83: 285–96.

———. 1991b. A multidimensional perspective on students' evaluations of teaching effectiveness: reply to Abrami and d'Apollonia (1991). *Journal of Educational Psychology* 83: 416–21.

———. 1994. Weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology* 86, no. 4: 631–48.

Marsh, H. W., and Dunkin, M. J. 1992. Students' evaluations of university teaching: a multidimensional perspective. In *Higher Education: handbook of Theory and Research,* ed. John C. Smart, 143–233. New York: Agathon Press.

Marsh, H.W., and D. Hocevar. 1991. The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education* 7, no. 1: 9–18.

Marsh, H.W., and L.A. Roche. 1992. The use of student evaluations of university instructors in different settings. *Australian Journal of Education* 36: 278–300.

———. 1993. The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal* 30, no. 1: 217–51.

———. 1997. Making students' evaluations of teaching effectiveness effective: the critical issues of validity, bias, and utility. *Journal of Educational Psychology* 52, no. 11: 1187–97.

———. 1999. Rely upon SET research. *American Psychologist* 54, no. 7: 517–8

Ory, J.C., and L.A. Braskamp. 1981. Faculty perceptions of the quality and usefulness of three types of evaluative information. *Research in Higher Education* 15, no. 3: 271–82.

Ryan, J.M., and P.D. Harrison. 1995. The relationship between individual instructional characteristics and the overall assessment of teaching effectiveness across different instructional contexts. *Research in Higher Education* 36, no. 5: 213–28.

Seldin, P. 1991. *The teaching portfolio.* Boston, MA: Anker Publishing.