9. Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness

Herbert W. Marsh*
Oxford University
herb.marsh@edstud.ox.ac.uk

Abstract

Students' evaluations of teaching effectiveness (SETs) have been the topic of considerable interest and a great deal of research in North America and, increasingly, universities all over the world. Research reviewed here indicated that SETs are:

- multidimensional;
- reliable and stable;
- primarily a function of the instructor who teaches a course rather than the course that is taught;
- relatively valid against a variety of indicators of effective teaching;
- relatively unaffected by a variety of variables hypothesized as potential biases; and
- Seen to be useful by faculty as feedback about their teaching, by students for use in course selection, and by administrators for use in personnel decisions

Key Words: Teaching effectiveness; reliability; construct validity; multidimensionality; bias; feedback interventions; longterm stability; profile analysis

Students' evaluations of teaching effectiveness (SETs) are commonly collected in U.S. and Canadian universities (Centra, 2003), are increasingly being used in universities throughout the world (e.g., Marsh &

*This chapter is a substantially revised version of the much longer chapter by Marsh and Dunkin (1997; also see Marsh 1984, 1987). I would like to thank particularly co-authors of earlier studies summarized in this review and colleagues who have offered suggestions on this and on my previous reviews of SET research. Requests for further information about this investigation should be sent to Professor Herbert W. Marsh, Department of Educational Studies, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY UK; E-mail: *herb.marsh@edstud.ox.ac.uk*.

R.P. Perry and J.C. Smart (eds.), The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective, 319–383.

Roche, 1997; Watkins, 1994), are widely endorsed by teachers, students, and administrators, and have stimulated much research spanning nearly a century. Numerous studies have related SETs to a variety of outcome measures broadly accepted by classroom teachers (e.g., learning inferred from classroom and standardized tests, student motivation, plans to pursue and apply the subject, positive affect, experimental manipulations of specific components of teaching, ratings by former students, classroom observations by trained external observers, and even teacher self-evaluations of their own teaching effectiveness). Considered here are the purposes for collecting SETs, SET dimensions, issues of reliability, validity and generalizability, potential biases in SETs, and the use of SETs for improving teaching effectiveness. As literally thousands of papers have been written, a comprehensive review is beyond the scope of this chapter. The reader is referred to reviews by: Aleamoni (1981); Braskamp, Brandenburg, and Ory (1985); Braskamp and Ory (1994); Cashin (1988); Centra (1979, 1989, 1993); Cohen, (1980, 1981); Costin, Greenough and Menges (1971); de Wolf (1974); Doyle (1975; 1983); Feldman (1976a, 1976b, 1977, 1978, 1979, 1983, 1984, 1986, 1987, 1988, 1989a, 1989b, 1992, 1993); Kulik and McKeachie (1975); Marsh (1982b, 1984, 1985, 1987); Marsh and Dunkin (1992, 1997); Marsh and Dunkin (1997, 2000); McKeachie (1963, 1973, 1979); Murray (1980); Overall and Marsh (1982); Remmers (1963); and Rindermann (1996).

PURPOSES FOR COLLECTING SETs

SETs are collected variously to provide:

- diagnostic feedback to faculty for improving teaching;
- a measure of teaching effectiveness for personnel decisions;
- information for students for the selection of courses and instructors;
- one component in national and international quality assurance exercises, designed to monitor the quality of teaching and learning; and
- an outcome or a process description for research on teaching (e.g., studies designed to improve teaching effectiveness and student outcomes, effects associated with different styles of teaching, perspectives of former students).

The first purpose is nearly universal, but the next three are not. Systematic student input is required before faculty are even considered

for promotion at many universities, but not at all at some others. At a few universities, students buy summaries of SETs in bookstores for purposes of course selection, but they are provided no access to the ratings in many other universities. The publication of SETs is controversial (Babad, Darley, & Kaplowitz, 1999; Perry, Abrami, Leventhal, & Check, 1979) and, not surprisingly, is viewed more positively by students than by teachers (Howell & Symbaluk, 2001). The existence of a program of students' evaluations of teaching is typically considered as one requirement of a good university in quality assurance exercises. Surprisingly, SET research has not been systematically incorporated into broader studies of teaching and learning (see Marsh & Dunkin, 1997).

DIMENSIONS OF SETs

Researchers and practitioners (e.g., Abrami & d'Apollonia, 1991; Cashin & Downey, 1992; Feldman, 1997; Marsh & Roche, 1993) agree that teaching is a complex activity with multiple interrelated components (e.g., clarity, interaction, organization, enthusiasm, feedback). Hence, it should not be surprising that SETs—like the teaching they are intended to represent—are also multidimensional. Particularly formative/diagnostic feedback intended to be useful for improving teaching should reflect this multidimensionality (e.g., a teacher can be organized but lack enthusiasm).

SET instruments differ in the quality of items, the way the teaching effectiveness construct is operationalized, and the particular dimensions that are included. The validity and usefulness of SET information depends upon the content and coverage of the items and the SET factors that they reflect. Poorly worded or inappropriate items will not provide useful information, while scores averaged across an ill-defined assortment of items offer no basis for knowing what is being measured. In practice, most instruments are based on a mixture of logical and pragmatic considerations, occasionally including some psychometric evidence such as reliability or factor analysis (Marsh & Dunkin, 1997). Valid measurement, however, requires a continual interplay between theory, research and practice. Careful attention should therefore be given to the components of teaching effectiveness that are to be measured. Whereas the usefulness of a SET program depends on more than having a well-designed instrument, this is an important starting point. Several theoretically defensible instruments with a well-defined factor structure have been reviewed (see Centra, 1993; Marsh 1987;

Marsh & Dunkin, 1997), but few have been evaluated extensively in terms of potential biases, validity, and usefulness of feedback.

Identifying the Dimensions to be Measured

Marsh and Dunkin (1997) noted three overlapping approaches to the identification, construction and evaluation of multiple dimensions in SET instruments: (1) empirical approaches such as factor analysis and multitrait-multimethod (MTMM) analysis; (2) logical analyses of the content of effective teaching and the purposes the ratings are intended to serve, supplemented by reviews of previous research and feedback from students and instructors (see Feldman, 1976b; also see Table 1); and (3) a theory of teaching and learning. In practice, most instruments are based on either of the first two approaches—particularly the second. The SET literature contains examples of instruments that have a well-defined factor structure, such as the four instruments presented by Marsh (1987; also see Centra, 1993; Jackson, Teal, Raines, Nansel, Force, Burdsal, 1999; Marsh & Dunkin, 1997; Richardson, 2005). Factor analyses have identified the factors that each of these instruments is intended to measure, demonstrating that SETs do measure distinct components of teaching effectiveness. The systematic approach used in the development of these instruments, and the similarity of the factors that they measure, supports their construct validity.

An important, unresolved controversy is whether the SET instruments measure effective teaching or merely behaviors or teaching styles that are typically correlated with effective teaching. In particular, is a teacher necessarily a poor teacher if he/she does not use higher order questions, does not give assignments back quickly, does not give summaries of the material to be covered, etc. (For further discussion, see McKeachie 1997; Scriven, 1981). Unless SETs are taken to be the criterion of good teaching, then it may be inappropriate to claim that a poor rating on one or more of the SET factors necessarily reflects poor teaching. Indeed, an often-cited complaint of SETs is that their use militates against some forms of effective teaching (see discussion on biases). Nevertheless, there is little or no systematic evidence to indicate that any of the typical SET factors is negatively related to measures of effective teachings (see discussion on validity). Furthermore, taken to its extreme, this argument could be used to argue against the validity of the type of behaviors that Scriven advocates should be measured by SETs or any other measure of effective teaching. Because teaching

Table 1: Categories of Effective Teaching Adapted From Feldman (1976b) and the Students' Evaluations of Educational Quality (SEEQ) and Endeavor factors Most Closely Related to Each Category

Feldman's Categories	SEEQ Factors
1) Stimulation of interest (I)	Instructor Enthusiasm
2) Enthusiasm (I)	Instructor Enthusiasm
3) Subject knowledge (I)	Breadth of Coverage
4) Intellectual expansiveness (I)	Breadth of Coverage
5) Preparation and organisation (I)	Organisation/Clarity
6) Clarity and understandableness (I)	Organisation/Clarity
7) Elocutionary skills (I)	None
8) Sensitivity to class progress (I/II)	None
9) Clarity of objectives (III)	Organisation/Clarity
10) Value of course materials (III)	Assignments/Readings
11) Supplementary materials (III)	Assignments/Readings
12) Perceived outcome/impact (III)	Learning/Value
13) Fairness, impartiality (III)	Examinations/Grading
14) Classroom management (III)	None
15) Feedback to students (III)	Examinations/Grading
16) Class discussion (II)	Group Interaction
17) Intellectual challenge (II)	Learning/Value
18) Respect for students (II)	Individual Rapport
19) Availability/helpfulness (II)	Individual Rapport
20) Difficulty/workload (III)	Workload/Difficulty

Note. The actual categories used by Feldman in different studies (e.g., Feldman, 1976, 1983, 1984) varied somewhat. Feldman (1976b) also proposed three higher-order clusters of categories, which are identified by I (presentation), II (facilitation), and III (regulation) in parentheses following each category.

effectiveness is a hypothetical construct, there is no measure (SETs or any other indicators) that IS effective teaching—only measures that are consistently correlated with a variety of indicators of teaching effectiveness.

THE STUDENTS' EVALUATION OF EDUCATIONAL QUALITY (SEEQ) INSTRUMENT

Strong support for the multidimensionality of SETs comes from research based on the SEEQ instrument (Marsh, 1982b; 1987; Marsh & Dunkin, 1997; Richardson, 2005). SEEQ measures nine factors (See Table 1). In the development of SEEQ, a large item pool was obtained

from a literature review, from forms in current usage, and interviews with faculty and students about what they saw as effective teaching. Students and teachers were asked to rate the importance of items; teachers were asked to judge the potential usefulness of the items as a basis for feedback, and open-ended student comments were examined to determine if important aspects had been excluded. These criteria, along with psychometric properties, were used to select items and revise subsequent versions, thus supporting the content validity of SEEQ responses. Marsh and Dunkin (1992, 1997; Marsh & Roche, 1994) also demonstrated that the content of SEEQ factors is consistent with general principles of teaching and learning, with particular emphasis on theory and research in adult education that is most relevant to higher education settings. As noted by Richardson (2005), the SEEQ instrument continues to be the most widely used instrument in published research. In summary, there is a strong empirical, conceptual, and theoretical basis for the SEEQ factors.

Factor analytic support for the SEEQ scales is particularly strong. The factor structure of SEEQ has been replicated in many published studies, but the most compelling support is provided by Marsh and Hocevar (1991a). Starting with an archive of 50,000 sets of class-average ratings (reflecting responses to 1 million SEEQ surveys), they defined 21 groups of classes that differed in terms of course level (undergraduate/graduate), instructor rank (teaching assistant/regular faculty), and academic discipline. The 9 a priori SEEQ factors were identified in each of 21 separate factor analyses. The average correlation between factor scores based on each separate analysis and factor scores based on the total sample was over .99. Whereas most SEEQ research has focused on student responses to the instrument, the same nine factors were identified in several large-scale studies of teacher self-evaluations of their own teaching using the SEEQ instrument (Marsh, Overall, & Kesler, 1979b; Marsh, 1983; also see Marsh, 1987, p. 295).

Studies using the "applicability paradigm" (see reviews by Marsh, 1986; Marsh & Roche, 1992; 1994; Watkins, 1994) in different Australian and New Zealand universities, in a cross-section of Australian Technical and Further Education institutions, and universities from a variety of different countries (e.g., Spain, Papua New Guinea, India, Nepal, Nigeria, the Philippines, and Hong Kong) provide support for the applicability of the distinct SEEQ factors outside the North American context in which they were developed. Watkins (1994) critically evaluated this research in relation to criteria derived from cross-cultural psychology. He adopted an "etic" approach to

cross-cultural comparisons that seeks to evaluate what are hypothesized to be universal constructs based on the SEEQ factors. Based on his evaluation of the applicability paradigm, Watkins (1994, p. 262) concluded, "the results are certainly generally encouraging regarding the range of university settings for which the questionnaires and the underlying model of teaching effectiveness investigated here may be appropriate."

OLDER, EXPLORATORY AND NEWER, CONFIRMATORY APPROACHES TO FACTOR ANALYSIS

Confirmatory factor analysis (CFA) has largely superseded traditional applications of exploratory factor analysis (EFA), and this has created an interesting disjuncture between SET research based on older instruments, derived from EFA and newer studies based on CFA (see related discussion by Abrami, d'Apollonia, & Rosenfield, 1993; 1997; Jackson et al., 1999; Marsh, 1987; 1991a; 1991b; Marsh & Dunkin, 1997; Toland & De Ayala, 2005). This is an important issue, because different practices in the application of EFA and CFA may give the appearance of inconsistent results if not scrutinized carefully (e.g., Toland & De Ayala, 2005). Given the extensive EFA evidence for SEEQ having a clearly defined, replicable structure, why would CFA provide apparently conflicting results?

The resolution of this dilemma is that the CFAs are typically based on a highly restrictive "independent clusters" model in which each item is allowed to load on one and only one factor, whereas exploratory factor analysis allows each item to cross-load on other factors. The exclusion of significant non-zero cross-loadings in CFA not only results in a poor fit to the data, but also distorts the observed pattern of relations among the factors. Although there are advantages in having "pure" items that load on a single factor, this is clearly not a requirement of a well-defined, useful factor structure, nor even a requirement of traditional definitions of "simple structure". The extensive EFA results summarized here clearly demonstrate that the SEEQ factor structure is well-defined, replicable over a diversity of settings, and stable over time, whereas the independent cluster model (e.g., Toland & De Ayala, 2005) does not provide an appropriate representation of the factor structure. In addressing this issue, Marsh (1991a, 1991b) also noted that an independent cluster model did not provide an adequate fit to the data, as many items had minor cross-loading on other factors. He randomly divided a large sample of classes into

groups, used empirical techniques to determine additional parameters, and then showed that this post hoc solution cross-validated well with the second sample. Thus, the existence of an a priori model based on CFA is the key to resolving the apparent anomaly identified by Toland and De Ayala.

An alternative solution to this problem is illustrated by Jackson et al. (1999), who compared CFA and EFA solutions based on analyses of a new set of 7,000 university classes from the Student's Perceptions of Teaching Effectiveness. This is an older instrument that has a well-established multidimensional structure with factors similar to those of SEEQ. Jackson et al. tested the replicability of an EFA solution based on previous results with a CFA based on new data, but allowed minor loadings for items with moderate cross-loadings in the original EFA. This a priori factor structure did not have an independent cluster solution, but the CFA model resulted in a good fit to the data and cross-validated well with EFAs based on both the new and the old data sets.

In summary, CFA offers important advantages over older, EFA approaches, but researchers must use care to evaluate appropriate models that accurately reflect factor structures and relations among variables. Whereas factor analytic research with appropriately designed instruments clearly supports a multidimensional perspective (e.g., the nine-factor solution for SEEQ), a more critical question is whether there is support for the discriminant validity and usefulness of the multiple factors in other research, such as studies evaluating relations with validity criteria, potential biases, and the usefulness of SETs for the purposes of improving teaching effectiveness.

Logical Approaches to the Identification of Dimensions of Teaching

Feldman (1976b; also see Table 1) logically derived a comprehensive set of components of effective teaching by categorising the characteristics of the superior university teacher from the student's point of view. He reviewed research that either asked students to specify these characteristics or inferred them on the basis of correlations with global SETs. In a content analysis of factors identified in well-defined multidimensional SET instruments, Marsh (1987) demonstrated that Feldman's categories tended to be more narrowly defined constructs than the empirical factors identified in many instruments—including SEEQ. Whereas SEEQ provided a more comprehensive coverage of Feldman's categories

than other SET instruments considered, most SEEQ factors represented more than one of Feldman's categories (e.g., Feldman's categories "stimulation of interest" and "enthusiasm" were both included in the SEEQ "instructor enthusiasm" factor). Surprisingly, there seems to have been no attempt to design and rigorously test an instrument based on Feldman's theoretical model of the components of effective teaching (but see Abrami et al., 1997).

GLOBAL SET RATINGS

Global or "overall" ratings cannot adequately represent the multidimensionality of teaching. They may also be more susceptible to context, mood and other potential biases than specific items that are more closely tied to actual teaching behaviors, leading Frey (1978) to argue that they should be excluded. In the ongoing debate on the value of global ratings, Abrami & d'Apollonia (1991; Abrami, d'Apollonia, & Rosenfield, 1997) seemed to initially prefer the sole use of global ratings for personnel decisions, whereas Marsh (1991b; Marsh & Bailey, 1993) preferred a profile of scores—including the different SEEQ factors, global ratings, expected grades, and prior subject interest ratings. In support of global ratings, Abrami et al argue the correlation between SETs and student learning in multisection validity studies is higher for global ratings than the *average* correlation based on specific rating factors. However, it is important to emphasize that student learning is systematically more highly correlated with specific components of SETs more logically related to SETs than to global SETs (see subsequent discussion of multi-section validity studies of student learning). Abrami et al. also argue that there exist a plethora of SET instruments that reflect a lack of clear consensus about the specific dimensions of SETs that are assessed in actual practice. However, it is also important to point out that Feldman (1976b) provided a comprehensive map of the specific SET dimensions that have been identified in empirical research that provides a basis for assessing those that are included on any particular instrument (see Table 1).

Although this debate continues, there is apparent agreement that an appropriately weighted average of specific SET factors may provide a workable compromise between these two positions. Along with other research exploring higher-order (more general) factors associated with SET dimensions (Abrami et al., 1997), this compromise acknowledges the underlying multidimensionality of SETs (Marsh & Roche, 1994). However, it also raises the thorny question of how

to weight the different SET components. Marsh and Roche (1994) suggested that for purposes of feedback to instructors (and perhaps for purposes of teacher input into personnel decisions), it might be useful to weight SET factors according to their importance in a specific teaching context as perceived by the teacher. Unresolved issues concerning the validity and utility of importance-weighted averages (e.g., Marsh, 1995), however, dictate caution in pursuing this suggestion.

Recent reviews of SET research (e.g., Apodaca & Grad, 2005; Hobson & Talbot, 2001) also noted that whereas there is general agreement on the appropriateness of a multidimensional perspective of SETs for purposes of formative feedback and instructional improvement, the debate about the most appropriate form of SET for summative purposes is unresolved: overall ratings, a multidimensional profile of specific SET factors, or global scores based on weighted or unweighted specific factors. Indeed, Marsh (1987; Marsh & Dunkin, 1997) recommended that teachers preparing a teaching portfolio for purposes of personnel decisions should be given the opportunity to use a multidimensional profile of SET scores to defend their approach to effective teaching—thereby implicitly endorsing use of a weighted-average approach.

In an attempt to discover how students weight different SET components in forming an overall evaluation, Ryan and Harrison (1995; Harrison, More & Ryan, 1996) conducted a policy-capturing experiment (also see Marsh & Groves, 1987) in which descriptions of hypothetical teachers were experimentally manipulated in relation to SEEQ factors. Results indicated that students demonstrated insight in forming overall SET ratings, using an appropriate weighting scheme that was consistent across students, thus supporting the use of a weighted-average approach based on weights derived from students.

Harrison, Douglas, and Burdsal (2004) specifically compared the usefulness of different strategies for obtaining global ratings (overall ratings, weighted averages with weights determined by students and teachers, unweighted averages, or higher-order factors based on higher-order factor analysis). Whereas they expressed a preference for a higher-order SET factor, they noted that results from all these approaches were highly correlated—suggesting that there was little empirical basis for choosing one over the others. However, conceptually and strategically there are apparently important differences that may affect the acceptability of SETs to academics, administrators, and students.

UNIT OF ANALYSIS PROBLEM

Misunderstanding about the appropriate unit of analysis continues to be a source of confusion and a critical methodological issue in SET research. Because of the nature of SETs, it is feasible to consider variation at the level of the individual student, the class or teacher, the department or faculty, or even an entire university. Fortunately, however, there is a clear consensus in SET research that the class-average or individual teacher is the appropriate unit of analysis, rather than the individual student (e.g., Cranton & Smith, 1990; Gilmore, Kane, & Naccarato, 1978; Howard & Maxwell, 1980; Marsh, 1987). As emphasized by Kane, Gillmore and Crooks (1976, p. 172), "it is the dependability of the class means, rather than the individual student ratings, that is of interest, and the class is the appropriate unit of analysis." Thus, support for the construct validity of student evaluation responses must be demonstrated at the class-average level (e.g., relations with class-average achievement, teacher self-evaluations), support for the factor structure of SETs should be based on a large, diverse set of class-average ratings, the reliability of responses is most appropriately determined from studies of interrater agreement among different students within the same course (also see Gilmore et al., 1978 for further discussion), and studies of potential bias (expected grades, class size, prior subject interest, workload/difficulty) should be based on class-average ratings.

Historically, due largely to limitations in statistical analysis available to them, SET researchers have had to choose a single unit of analysis. In such cases, the class-average is almost always the appropriate unit of analysis. However, as suggested by Marsh and Dunkin (1997; Marsh, 1987), advances in the application of multilevel modeling open up new opportunities for researchers to simultaneously consider more than one unit of analysis (e.g., individual student and class) within the same analysis.

Although commercial packages have greatly facilitated the application of multilevel modeling, there are only a few examples of multilevel modeling in SET research (e.g., Marsh and Hattie, 2002; Marsh, Rowe, and Martin, 2002; Ting, 2000; Toland & De Ayala, 2005; Wendorf & Alexander, 2004). It is important to emphasize that the typical analysis of class-average SETs is not invalidated by the existence of a multilevel structure to the data, in which there is significant variation at both the individual student and class levels, but this multilevel structure does invalidate most analyses conducted at the

individual student level. More importantly, a systematic evaluation of the multilevel structure of the data allows researchers to pursue new questions not adequately addressed by conventional analyses. Thus, for example, whereas researchers have routinely evaluated variance components associated with individual students and classes, a more complete analysis of the multilevel structure might address, for example, how SETs vary from department to department and the characteristics of departments associated with this variation, or even differences between entire universities (Marsh, Rowe, and Martin, 2002). In the near future it is likely that multilevel modeling will become widely used in SET research.

Implicit Theories and the Systematic Distortion Hypothesis

Theoretical work on the implicit theories that people use to make ratings and the systematic distortion hypothesis based largely on personality research (e.g., Cronbach, 1958) has been applied to SET research to provide an alternative explanation for the robustness of factor structures based on a well-designed, multidimensional SET instrument. Marsh (1984; also see Marsh, 1987) noted, for example, that if a student's implicit theory of behavioral covariation suggests that the occurrences of behaviors X and Y are correlated and if the student rates the teacher high on X, then the teacher may also be assumed to be high on Y, even though the student has not observed Y. The systematic distortion hypothesis predicts that traits can be rated as correlated (based on implicit theories), whereas actual behaviors reflecting these traits are not correlated.

In a study particularly relevant to implicit theories, Cadwell and Jenkins (1985) specifically noted the factor analytic research based on SEEQ was "particularly impressive" (p. 383), but suggested that the strong support for the factor structure was due to semantic similarities in the items. To test this speculation, they asked student to make ratings of teaching effectiveness based on scripted scenarios (sets of 8 one-sentence descriptions depicting the presence or absence of each behavior) derived from various combinations of SEEQ items. However, in their critique of the Cadwell and Jenkins (1985) study, Marsh & Groves (1987) noted many methodological problems and conceptual ambiguities; thus, interpretations should be made cautiously. In particular, students were given inadequate or conflicting information that required them to rely on implicit theories and re-interpretations

of the meaning of the behaviors to make sense of the task. For example, students were told whether or not an instructor "often summarized material in a manner that aided comprehension" (p. 386) and "presented a brief overview of the lecture content" (p. 386), as a basis for responding to the SEEQ item "the objectives of the course were clearly stated and pursued", but were given no information about the actual pursuit of course objectives. Even more problematic, students were asked to make ratings on the basis of apparently contradictory behavioral descriptions. For example, they were told that the same instructor "summarized material in a manner that aided comprehension" (p. 386) but did *not* "present a brief overview of the lecture content" (p. 386). Hence, students in this study were forced to make inferences about SEEQ items based on the information available or to devise plausible explanations for apparently contradictory information, to make sense of the task. Marsh and Grove argued that these and other conceptual and methodological problems precluded any justifiable conclusions about the effect of semantic similarities and implicit theories. Nevertheless, Cadwell and Jenkins did find that most of the systematic variation in responses to each SEEQ item was associated with differences in the experimentally manipulated teaching behaviors designed to parallel that item, thus supporting the construct validity of SEEQ responses.

More recently, Renaud and Murray (2005) conducted one of the most detailed tests of the systematic distortion hypothesis in relation to implicit theories. Noting the failure of most previous research, such as the Caldwell and Jenkins (1985) study, to include behaviors based on actual classrooms, they considered: (a) student ratings of teaching effectiveness (SETs) under typical conditions for a sample of 32 teachers; (b) frequency counts of observable teaching behaviors based on videotapes of these same teachers; and (c) ratings of the conceptual similarity of all possible pairs of items used in these tasks. In support of the validity of students' implicit theories, covariation between SET items was substantially related to covariation among teaching behaviors. However, covariation between SETs and similarity ratings was somewhat higher, suggesting the possibility of a semantic distortion in addition to covariation among ratings consistent with actual behaviors. However, whereas the application of implicit theories to SET research has been heuristic, apparently inherent complexities and difficult methodological problems like those discussed by Marsh & Groves (1987) and by Renaud and Murray (2005) mean that unambiguous interpretations are unlikely to result from these studies.

Summary of the Dimensionality of SETs

Many SET instruments are not developed using a theory of teaching and learning, a systematic logical approach that ensures content validity, or empirical techniques such as factor analysis. Appropriately constructed SET instruments and particularly research based on SEEQ provide clear support for the multidimensionality of the SET construct. Whereas some instruments based on far fewer items provide evidence of fewer factors, it is clear that students are able to differentiate between distinct components of effective teaching. Indeed, the classification scheme developed by Feldman (1987; see Table 1) provides an appropriate framework for evaluating the comprehensiveness of any particular instrument. The debate about which specific components of teaching effectiveness can and should be measured has not been resolved, although there seems to be consistency in those identified in response to the most carefully designed instruments such as SEEQ, which are apparently applicable to a wide diversity of educational settings. Furthermore, it is important to note that many poorly constructed student evaluation surveys fail to provide a comprehensive multidimensional evaluation, thus undermining their usefulness, particularly for diagnostic feedback. "Home-made" SET surveys constructed by lecturers themselves, or by committees, are particularly susceptible to such deficiencies, and compounded by the likelihood that aspects of teaching excluded from the survey are those which tend to be the most neglected in practice. Such "one shot" instruments are rarely evaluated in relation to rigorous psychometric considerations and revised accordingly. SET instruments should be designed to measure separate components of teaching effectiveness, and support for both the content and the construct validity of the multiple dimensions should be evaluated.

Reliability, Stability, Generalizability, and Applicability

Reliability

Traditionally, reliability is defined on the basis of the extent of agreement among multiple items designed to measure the same underlying construct, using indexes such as coefficient alpha. This approach, although potentially useful, does not provide an adequate basis for assessing the reliability of SET responses. The main source of variability is lack of agreement among different students' ratings of the same

teacher rather than lack of agreement among different items. Hence, the reliability of SETs is most appropriately determined from studies of interrater agreement that assess lack of agreement among different students within the same course (see Gilmore et al., 1978 for further discussion). The correlation between responses by any two students in the same class (i.e., the single rater reliability; Marsh, 1987) is typically in the .20s but the reliability of the *class-average* response depends upon the number of students rating the class: .95 for 50 students, .90 for 25 students, .74 for 10 students, and .60 for five students. Given a sufficient number of students, the reliability of class-average SETs compares favourably with that of the best objective tests.

Although there are more sophisticated approaches to error that can incorporate both lack of agreement among items and students as well as other sources of error, such generalizability research typically shows that lack of agreement among students is by far the largest source of error (see Gilmore et al., 1978 for further discussion). In these analyses, differences between responses by individual students are typically considered to reflect random measurement error. More recent developments of multilevel modeling allow researchers to simultaneously incorporate both the class and the individual student into the same analysis. This would allow researchers to determine, for example, individual student characteristics that may explain variation among students nested within classes, how these individual characteristics might affect class-average ratings, and how these might interact with class-level characteristics to influence class-average ratings.

STABILITY

Sadly, there is a broad range of cross-sectional and longitudinal research demonstrating that teaching effectiveness—no matter how measured—tends to decline with age and years of teaching experience (see reviews by Marsh, 1987; Marsh & Dunkin, 1997). At best, there is limited evidence of an increase in teaching effectiveness over the first few years of teaching, followed by a gradual decline in teaching effectiveness. Hence, it is not surprising that cross-sectional studies typically report that SETs are also negatively related to age and years of teaching experience (Feldman, 1983; Renaud & Murray, 1996), although there is some suggestion that SETs may increase slightly during the first few years of teaching (Marsh & Dunkin, 1997). Also, this effect may vary somewhat with the particular SET dimension. Furthermore, these results are typically based on average responses aggregated across many

teachers so that, perhaps, there are large individual differences for particular teachers—some improving and others declining—that are lost when averaged across teachers. Cross-sectional studies provide a poor basis for inferring how ratings of the same person will change over time.

In a true longitudinal study, Marsh and Hocevar (1991b) examined changes in ratings of a diverse sample of 195 teachers who had been evaluated continuously over a 13-year period. Based on an average of more than 30 sets of ratings for each teacher, they found that the mean ratings for their cohort of 195 teachers showed almost no systematic changes in any of the SEEQ factors for the total group or for subsamples with little, intermediate, or substantial amounts of teaching experience at the start of the 13-year longitudinal study. Furthermore, whereas there were some individual differences in this trend, there was only a small number of teachers who showed systematic increases or decreases over time. Although it is discouraging that the feedback from the ratings alone did not lead to systematic improvement, it is encouraging that this group of teachers who had received so much SET feedback did not show the systematic declines in teaching effectiveness that appear to be the norm (also see Kember, Leung, & Kwan, 2002). The Marsh and Hocevar study is particularly important in showing the stability of the SEEQ factor structure over time and the stability of SETs over an extended period of time.

Generalizability

Student versus alumni ratings. Some critics suggest that students cannot recognize effective teaching until being called upon to apply their mastery in further coursework or after graduation. However, cross-sectional studies show good agreement between responses by current students and alumni (see Marsh, 1987; Centra, 1979, 1989). In a true longitudinal study (Overall & Marsh, 1980), ratings in 100 classes correlated .83 with ratings by the *same* students when they again evaluated the same classes retrospectively several years later, at least one year after graduation. These studies demonstrate that SETs for alumni and current students are very similar.

Teacher versus course effects. Researchers have also explored the correlation of SETs in different courses taught by the same instructor or in the same course taught by different teachers. Results (Marsh, 1987; Marsh & Dunkin, 1997; also see Rindermann & Schofield, 2001) demonstrate that SETs are primarily due to the instructor who teaches

a class and not the particular class being taught. Thus, for example, Marsh (1987, p. 278) reported that for the overall instructor rating, the correlation between ratings of different instructors teaching the same course (i.e., a course effect) was −.05, while correlations for the same instructor in different courses (.61) and in two different offerings of the same course (.72) were much larger. These results support the validity of SETs as a measure of teacher effectiveness, but not as a measure of course effectiveness independent of the teacher.

This research on teacher and course effects also has important implications for the compilation of normative archives used to assess teaching effectiveness, based on ratings of the same teacher over time in different courses. Gilmore, Kane, and Naccarato (1978), applying generalizability theory to SETs, suggested that ratings for a given instructor should be averaged across different courses to enhance generalizability. If it is likely that an instructor will teach many different classes during his or her subsequent career, then tenure decisions should be based upon as many different courses as possible—Gilmore, Kane, and Naccarato, suggest at least five. These recommendations require that a longitudinal archive of SETs is maintained for personnel decisions. These data would provide the basis for more generalizable summaries, the assessment of changes over time, and the determination of which particular courses are best taught by a specific instructor. Indeed, the evaluation of systematic change in SETs of the same teacher over time would also provide an alternative basis of comparison that was not based on how ratings of a given teacher compared with those by other teachers. It is most unfortunate that some universities systematically collect SETs, but fail to keep a longitudinal archive of the results.

GENERALIZABILITY OF PROFILES

Marsh and Bailey (1993) used multivariate profile analysis to demonstrate that each teacher has a characteristic profile on the 9 SEEQ scores (e.g., high on organisation and low on enthusiasm). For each teacher who had been evaluated continuously over 13 years, Marsh and Bailey determined a characteristic profile of SEEQ factors based on all the SETs of each teacher. Each teacher's characteristic profile was distinct from the profiles of other teachers, generalised across course offerings over the 13-year period, and even generalised across undergraduate and graduate level courses. Indeed, the generalizability of the profile of SEEQ scores was as strong as or stronger than the generalizability of the individual SEEQ factors and global ratings over time. Similarly,

Hativa (1996) also demonstrated that SETs were highly stable in terms both of the level and profile based on multiple ratings of the same teachers teaching the same course on multiple occasions. These results provide further support for the multidimensionality of SETs and their generalizability.

This support for the existence of teacher-specific profiles also has important implications for the use of SETs as feedback and for the relation of SETs to other criteria such as student learning. For example, presentation of an appropriate profile of SET factors (Marsh, 1987) provides clear evidence about relative strengths and weaknesses in teaching effectiveness. Given this stability of profiles, Marsh and Bailey lament that so little research has evaluated how specific profiles of SETs are related to student learning, other validity criteria, potentially biasing factors, and other correlates of SETs. For example, meta-analyses show that SETs are related to student learning and feedback interventions, and that the effect sizes vary systematically and logically with the specific SET component. However, there has been almost no research to establish how characteristic profiles are related to these criteria. Thus, for example, a profile in which both enthusiasm and organization are high might be particularly conducive to learning—beyond what can be explained in terms of either of these SET factors considered in isolation.

Student Written Comments—Generality Across Different Response Forms

Braskamp and his colleagues (Braskamp et al., 1985; Braskamp, Ory, & Pieper, 1981; Ory, Braskamp & Pieper, 1980) examined the usefulness of students' written comments and their relation to SET rating items. Student comments were scored for overall favorability with reasonable reliability and these overall scores correlated with responses to the overall rating item ($r = .93$), close to the limits of the reliability of the two indicators (Ory, Braskamp & Pieper, 1980). Braskamp, Ory, & Pieper (1981) sorted student comments into one of 22 content categories and evaluated comments in terms of favorability. Comment favorability was again highly correlated with the overall instructor rating (.75).

In a related study, Ory and Braskamp (1981) simulated results about a hypothetical instructor, consisting of written comments in their original unedited form and rating items—both global and specific. The rating items were judged as easier to interpret and more comprehensive for both personnel decisions and self-improvement, but other

aspects of the written comments were judged to be more useful for purposes of self-improvement. Speculating on these results, the authors suggested that "the nonstandardized, unique, personal written comments by students are perceived as too subjective for important personnel decisions. However, this highly idiosyncratic information about a particular course is viewed as useful diagnostic information for making course changes" (pp. 280–281). However, Murray (1987) reported that for purposes of feedback, teachers more strongly endorsed ratings of specific components of teaching effectiveness (78%) than written comments (65%), although global ratings were seen as even less useful (54%).

Lin, McKeachie, and Tucker (1984) reported that the impact of statistical summaries based on specific components of SETs was enhanced by written comments for purposes of promotional decisions—although the effects of research productivity were much larger. However, because they did not consider comments alone, or comments that were inconsistent with the statistical summaries in their experimental simulation study, there was no basis for comparing the relative impact of the two sources of information. Perhaps, because student comments are not easily summarized (due to the effort required as well as their idiosyncratic nature, which is dependent upon the specific class context), it may be more appropriate simply to return written comments to teachers along with appropriate summaries of the SET ratings. A useful direction for further research would be to evaluate more systematically whether this lengthy and time consuming exercise provides useful and reliable information that is not obtainable from the more cost effective use of appropriate multidimensional rating items. Unfortunately, there has apparently been no research to compare results of multidimensional content categories based on written comments with a well-defined multidimensional profile of SET ratings to evaluate the convergent and discriminant validity of both sources of information.

VALIDITY

THE CONSTRUCT VALIDATION APPROACH TO VALIDITY

SETs, as one measure of teaching effectiveness, are difficult to validate, since no single criterion of effective teaching is sufficient. Historically, researchers have emphasised a narrow, criterion-related approach to

validity in which student learning is the only criterion of effective teaching. This limited framework, however, inhibits a better understanding of what is being measured by SETs, of what can be inferred from SETs, and how findings from diverse studies can be understood within a common framework. Instead, Marsh (1987) advocated a construct validation approach in which SETs are posited to be positively related to a wide variety of other indicators of effective teaching and specific rating factors are posited to be most highly correlated with variables to which they are most logically and theoretically related. Although student learning—perhaps inferred in a variety of different ways—is clearly an important criterion of effective teaching, it should not be the only criterion to be considered. Hence, within this broader framework, evidence for the long-term stability of SETs, the generalizability of ratings of the same instructor in different courses, and the agreement in ratings of current students and alumni can be interpreted as support for the validity of SETs.

The most widely accepted criterion of effective teaching, appropriately, is student learning. However, other criteria include changes in student behaviors, instructor self-evaluations, ratings by colleagues and administrators, the frequency of occurrence of specific behaviors observed by trained observers, and experimental manipulations. A construct validity approach to the study of SETs now appears to be widely accepted (e.g., Cashin, 1988; Howard, Conway, & Maxwell, 1985). A difficulty in this approach is obtaining criterion measures that are reliably measured and that validly reflect effective teaching. If alternative indicators of teaching effectiveness are not reliable and valid, then they should not be used as indicators of effective teaching for research, policy formation, feedback to faculty, or personnel decisions.

Student Learning—The Multisection Validity Study

The most widely accepted criterion of student learning is performance on standardized examinations. However, examination performance typically cannot be compared across different courses except in specialized settings. In order to address this issue, SET researchers have proposed the multisection validity paradigm in which it may be valid to compare teachers in terms of operationally defined learning, and to relate learning to SETs.

In the ideal multisection validity study (Cohen, 1981; Feldman, 1989b; Marsh, 1987; Sullivan & Skanes, 1974) there are many sections of a large multisection course; students are randomly assigned to sections so as to minimize initial differences between sections; pretest measures that correlate substantially with final course performance serve as covariates; each section is taught completely by a separate instructor; each section has the same course outline, textbooks, course objectives, and final examination; the final examination is constructed to reflect the common objectives and, if there is a subjective component, it is graded by an external person; students in each section evaluate teaching effectiveness on a standardized evaluation instrument, preferably before they know their final course grade and without knowing how performances in their section compare with those of students in other sections; and section-average SETs are related to section-average examination performance, after controlling for pretest measures.

Despite methodological problems (Abrami, d'Apollonia, & Cohen, 1990; Marsh & Dunkin, 1992, 1997; Marsh & Roche, 1994), meta-analyses of multisection validity research have supported the validity of the SETs by demonstrating that the sections that evaluate the teaching as most effective are also the sections that perform best on standardized final examinations (Cohen, 1981, 1987; Feldman, 1989b). Cohen (1987), in his summary of 41 "well-designed" studies, reported that the mean correlations between achievement and different SET components were Structure (.55), Interaction (.52), Skill (.50), Overall Course (.49), Overall Instructor (.45), Learning (.39), Rapport (.32), Evaluation (.30), Feedback (.28), Interest/Motivation (.15), and Difficulty (−.04), in which all but the last two were statistically significant. Feldman (1989b) extended this research by demonstrating that many of Cohen's broad categories were made up of more specific components of SETs that are differentially related to student achievement. Thus, for example, Cohen's broad "skill" category was represented by 3 dimensions in Feldman's analysis, which correlated with achievement .34 (instructor subject knowledge), .56 (clarity and understandableness), and .30 (sensitivity to class level and progress). Cohen (1987; also see Feldman, 1989b; 1990) also reported that correlations were higher when specific SET components were measured with multi-item scales instead of single items. This research demonstrates that teachers who receive better SETs are also the teachers from whom students learn the most. Perhaps more than any other area of SET research, results based on the multisection validity paradigm support the validity of SETs.

Evaluations of Teaching Effectiveness by Different Evaluators

Teaching effectiveness can be evaluated by current students, former students, the instructor him/herself, colleagues, administrators, or trained external observers.

Self-evaluations. Instructors can be asked to evaluate themselves in a wide variety of educational settings, even using the same instrument used by their students, so as to provide tests of convergent and divergent validity. Despite the apparent appeal of instructor self-evaluations as a criterion of effective teaching, it has had limited application. Feldman's (1989b) meta-analysis of correlations between SETS and self-evaluations, based on only 19 studies, reported a mean *r* of .29 for overall ratings and mean *r*s of .15 to .42 for specific SET components. Marsh (1982c, 1987; Marsh, Overall, & Kesler, 1979b) conducted two studies in which large numbers of instructors evaluated their own teaching on the same multifaceted evaluation instrument that was completed by students. In both studies: separate factor analyses of SETs and self-evaluations identified the same SEEQ factors; student-teacher agreement on every dimension was significant (median *r*s of .49 and .45) and typically larger than agreement on overall teaching effectiveness (*r*s of .32); mean differences between student and faculty responses were small and unsystematic. Particularly important for the multidimensional perspective of SETs, MTMM analyses provided support for both convergent and discriminant validity of the ratings. Hence, not only was there general student-teacher agreement on teaching effectiveness overall, the student-teacher agreement was specific to each of the different SET factors (e.g., organization, enthusiasm, rapport).

Peer evaluations. Colleague, peer, and administrator ratings that are *not* based upon classroom visitation are sometimes substantially correlated with SETS, but it is likely that colleague ratings are based on information from students (Marsh, 1987; Marsh & Dunkin, 1992, 1997; Marsh & Roche, 1994). In contrast, colleague and administrator ratings based on classroom visitation do not appear to be very reliable (i.e., ratings by different peers do not even agree with each other) or to correlate substantially with SETs or with any other indicator of effective teaching (see Marsh, 1987; Centra, 1979). While these findings neither support nor refute the validity of SETs, they suggest that the colleague and administrator ratings based on classroom visitation are not valid indicators of teacher effectiveness (also see Murray, 1980).

External observer ratings. Murray (1980) concluded that SETs "can be accurately predicted from external observer reports of specific classroom teaching behaviors" (1980, p. 31). For example, Cranton and Hillgartner (1981) examined relationships between SETs and specific teaching behaviors observed on videotaped lectures in a naturalistic setting; SETs of organisation were higher "when instructors spent time structuring classes and explaining relationships;" SETs of effectiveness of student-teacher interaction and discussion were higher "when professors praised student behavior, asked questions and clarified or elaborated student responses" (p. 73).

In one of the most ambitious observation studies, Murray (1983) trained observers to estimate the frequency of occurrence of specific teaching behaviors of 54 university instructors who had previously obtained high, medium or low SETs in other classes. A total of 18 to 24 sets of observer reports were collected for each instructor. The median of single-rater reliabilities (i.e., the correlation between two sets of observational reports) was .32, but the median reliability for the average response across the 18–24 reports for each instructor was .77. Factor analysis of the observations revealed nine factors, and their content resembled factors in SETs described earlier (e.g., clarity, enthusiasm, interaction, rapport, organisation). The observations significantly differentiated among the three criterion groups of instructors. Unfortunately, Murray only considered SETs on an overall instructor rating item, and these were based upon ratings from a previous course rather than the one that was observed. Hence, MTMM-type analyses could not be used to determine if specific observational factors were most highly correlated with matching student rating factors. The findings do show, however, that instructors who are rated differently by students do exhibit systematically different observable teaching behaviors, and provide clear support for SETs in relation to these specific behaviors.

Multiple evaluators with different perspectives. Howard, Conway, and Maxwell (1985; also see Feldman, 1989a and discussion of his review by Marsh and Dunkin, 1992, 1997) compared multiple indicators of teaching effectiveness for 43 target teachers who were each evaluated in one course by: current students in the course (mean $N = 34$ per class); former students who had previously taken the same or similar course taught by the target teacher (minimum $N = 5$); one colleague who was knowledgeable of the course content and who attended two class sessions taught by the target teacher; and 8 advanced graduate students specifically trained in judging

teaching effectiveness, who attended two class sessions taught by the target teacher. Howard et al. concluded that "former-students and student ratings evidence substantially greater validity coefficients of teaching effectiveness than do self-report, colleague and trained observer ratings" (p. 195). Whereas self-evaluations were modestly correlated with current SETs (.34) and former SETs (.31), colleague and observer ratings were not significantly correlated with each other, current SETs, or self-evaluations.

Experimentally Manipulated Teacher Behaviors

A limited amount of research has related SETs to experimentally manipulated teaching situations. Studies of teacher clarity and teacher expressiveness (see reviews by Marsh, 1987; Marsh & Dunkin, 1992, 1997; Marsh & Roche, 1994) demonstrate the important potential of this approach. Both these teaching behaviors are amenable to experimental and correlational designs, can be reliably judged by students and by external observers, are judged to be important components of teaching effectiveness by students and by teachers, and are related to student achievement in naturalistic and experimental studies. In experimental settings, scripted lessons which differ in these teaching behaviors are videotaped, and randomly assigned groups of subjects view different lectures, evaluate teaching effectiveness, and complete achievement tests. Manipulations of these specific behaviors are significantly related to SETs and substantially more strongly related to matching SET dimensions than to nonmatching SET dimensions. These results support the inclusion of clarity and expressiveness on SET instruments, demonstrate that SETs are sensitive to natural and experimentally manipulated differences in these teaching behaviors, and support the construct validity of the multidimensional SETs with respect to these teaching behaviors. More generally, the direct manipulation of teaching behaviors and the experimental control afforded by laboratory studies are an important complement to quasi-experimental and correlational field studies.

Summary and Implications of Validity Research

Effective teaching is a hypothetical construct for which there is no adequate single indicator. Hence, the validity of SETs or of any other indicator of effective teaching must be demonstrated through a construct validation approach. SETs are significantly and consistently

related to the ratings of former students, student achievement in multisection validity studies, faculty self-evaluations of their own teaching effectiveness, and, perhaps, the observations of trained observers on specific processes such as teacher clarity. This provides support for the construct validity of the ratings. In contrast, colleague and administrator ratings based on classroom visitation are not systematically related to SETs or other indicators of effective teaching, which calls into question their validity as measures of effective teaching.

Nearly all researchers argue that it is necessary to have multiple indicators of effective teaching whenever the evaluation of teaching effectiveness is to be used for personnel decisions. It is, however, critical that the validity of *all* indicators of teaching effectiveness, not just SETs, be systematically examined before they are actually used. The heavy reliance on SETs as the primary measure of teaching effectiveness stems in part from the lack of support for the validity of any other indicators of effective teaching. This lack of viable alternatives—rather than a bias in favor of SETs—seems to explain why SETs are used so much more widely than other indicators of effective teaching.

Whereas SET validity research has been dominated by a preoccupation with student achievement and the multisection validity paradigm, there is too little research relating SETs to other criteria. Thus, for example, Marsh (1987; Marsh & Dunkin, 1992, 1997; Marsh & Roche, 1994) discussed the validity of SETs in relation to student motivation, self-concept, affective criteria, subsequent coursework selection, student study strategies and the quality of student learning. Whereas he argued that it is imperative to expand the range of validity criteria in SET research substantially, this plea has apparently not been pursued in subsequent published research. There is also surprisingly little research validating SETs in relation to experimentally manipulated teaching situations, even though there are some good demonstrations of this approach based on teacher clarity and teacher expressiveness (see Marsh, 1987).

Practitioners and researchers alike give lip-service to the adage that teaching effectiveness should be evaluated with multiple indicators of teaching—not just SETs. To this prescription I would like to add the caveat that all indicators of teaching effectiveness for formative or summative assessment should be validated from a construct validity approach prior to being integrated into practice. However, there are few other indicators of teaching effectiveness whose use is systematically supported by research findings. As noted by Cashin (1988), "student ratings tend to be statistically reliable,

valid, and relatively free from bias, probably more so than any other data used for faculty evaluation" (p. 5).

RESEARCH PRODUCTIVITY: A TEACHING-RESEARCH NEXUS

Teaching and research are typically seen as the most important products of university academics. Marsh (1987; Marsh and Hattie, 2002) contrasted opposing theoretical perspectives positing that indicators of the two activities should be positively correlated, negatively correlated, or uncorrelated.

There is a clear rationale for a positive nexus of reciprocal relations between teaching and research. Teachers who are active researchers are more likely to be: on the cutting edge of their discipline; aware of international perspectives in their field; and convey a sense of excitement about their research and how it fits into a larger picture. The process of teaching forces academics to clarify the big picture into which their research specialization fits, clarifying their research and reinforcing research pursuits through sharing it with students. Indeed, without this positive relation between teaching and research, one basis for funding modern research universities to pursue research as well as providing teaching is undermined.

The case can also be made as to why teaching and research are incompatible. Blackburn (1974) noted, for example, that unsatisfactory classroom performance might result from academics neglecting their teaching responsibilities in order to pursue research. The time and energy required to pursue one is limited by the time demands of the other, whereas the motivation and reward structures that support the two activities might be antagonistic as well.

Hattie and Marsh (1996) conducted a comprehensive meta-analysis of the relation between teaching and research among University academics. Based on 58 articles contributing 498 correlations, the overall correlation was 0.06 (see also Feldman, 1987; Centra, 1983). They searched for mediators and moderators to this overall correlation, with little success. The overall conclusion of a zero relation was found across: disciplines, various measures of research output (e.g., quality, productivity, citations), various measures of teaching quality (student evaluation, peer ratings), and different categories of university (liberal, research). Based on this review they concluded that the common belief that research and teaching are inextricably entwined is an enduring myth. At best, research and teaching are loosely coupled.

Marsh and Hattie (2002) pursued suggestions from the literature to better understand this belief in a positive nexus between teaching and research, and to discover situations or characteristics that reinforce a positive teaching-research relation. Data were based on a representative sample of academics from one research university who had extensive data on teaching effectiveness (SETs), externally monitored research productivity over three years, and completed a detailed survey on teaching and research constructs (self-ratings of ability, satisfaction, personal goals, motivation, time spent, supporting activities, and beliefs in a nexus). They began by testing Marsh's (1984; 1987) theoretical model in which the near-zero relation between teaching and research outcomes is a function of the counterbalancing positive relation between teaching and research abilities and the negative relation between time required to be effective at teaching and research and, perhaps, the motivation to be a good researcher and a good teacher. They found limited support for theoretical predictions. Whereas there was a substantial negative relation between time spent on teaching and research and no significant relation between teaching and research outcomes, there were no statistically significant relations between teaching and research ability or between teaching and research motivation.

Consistently with predictions, teaching ability had a moderate effect on teaching effectiveness and research ability had a substantial effect on research publications. The corresponding motivation and time variables had no significant effect on the teaching and research outcome variables (beyond what can be explained in terms of ability). In support of the posited antagonism between teaching and research, research ability had positive effects on research motivation and time, but negative effects on teaching motivation and time. Teaching ability had no significant effect on teaching motivation or teaching time, but it had a negative effect on research motivation. However, there was no support for the fundamental assumption that the ability to be a good teacher and the ability to be a good researcher are positively related. Indeed, because self-ratings are likely to be positively biased by potential biases (e.g., halo effects), it was quite surprising that these self-rating variables were not positively correlated.

Marsh and Hattie (2002) explored further research and teaching variables that might mediate the relations between ability and outcomes, including the belief that there is a nexus—that teaching contributes to research, or vice versa. Academics who believed that research contributes to teaching had more research publications and

higher self-ratings of research. However, beliefs in this nexus had no relation to the corresponding measures of teaching. In contrast, the belief that teaching contributes to research was not significantly related to self-ratings or outcomes for either teaching or research. Using multilevel modeling techniques they found that the near-zero correlation between teaching and research was consistent across the 20 academic departments included in their research, suggesting that differences in departmental ethos (or any other departmental characteristic) apparently had little impact on the teaching-research relation. They also explored a wide variety of potential moderators of the teaching-research relation to predict those who were high in both, but these results were also non-significant and supported the generality of the near-zero correlation between teaching and research.

In summary, this research supports the notion of teaching and research as reasonably independent constructs. While these findings seem neither to support nor refute the validity of SETs, they do demonstrate that measures of research productivity cannot be used to infer teaching effectiveness or vice versa. However, this research program has also stimulated a fierce debate about its implications. Particularly in the UK, the findings have been interpreted to mean that research and teaching functions of universities should be separated, fuelling further outrage within an academic community whose beliefs of integration prevail. It is noted, however, that a zero correlation need not lead to this separation—it means that there are just as many good teachers and researchers, not so good teachers and researchers, good researchers and not so good teachers, and good teachers and not so good researchers—independence of association does not mean that the two are necessarily "separate" for all. For those who believe so fervently that there is a positive teaching-research nexus, the failure to demonstrate it is seen to reflect inappropriate research. My belief is that a positive teaching-research nexus should be a goal of universities (to increase the number of academics who are both good teachers *and* good researchers), but empirical research provides little evidence that universities have been successful in doing so.

POTENTIAL BIASES IN STUDENTS' EVALUATIONS

The voluminous literature on potential biases in SETs is frequently atheoretical, methodologically flawed, and not based on well-articulated operational definitions of bias, thus continuing to fuel (and to be fuelled by) myths about bias (Feldman, 1997; Marsh,

1987; Marsh & Dunkin, 1997). Marsh listed important methodological problems in this research including: (a) implying causation from correlation; (b) use of an inappropriate unit of analysis (the class-average is usually appropriate, whereas the individual student is rarely appropriate); (c) neglect of the multivariate nature of SETs and potential biases; (d) inappropriate operational definitions of bias and potential biasing factors; and (e) inappropriate experimental manipulations.

Proper evaluation of validity, utility, and potential bias issues in SETs (see Feldman, 1998; Marsh & Dunkin, 1992; Marsh & Roche, 1997) demands the rejection of such flawed research, including narrow criterion-related approaches to bias. Instead, as for validity research, I use a broad construct validity approach to the interpretation of bias, which recognizes that (a) effective teaching and SETs designed to measure it are multidimensional; (b) no single criterion of effective teaching is sufficient; and (c) theory, measurement, and interpretations of relations with multiple validity criteria and potential biases should be evaluated critically across different contexts and research paradigms. Recognition of the *multidimensionality* of teaching and of SETs is fundamental to the evaluation of competing interpretations of SET relations with other variables. Although a construct validity approach is now widely accepted in evaluating various aspects of validity, its potential usefulness for the examination of bias issues has generally been ignored.

Marsh and Dunkin (1997; also see Centra, 1979; Marsh, 1987; also see Table 2 for a summary of typical relations between SETs and potential biases, based on earlier reviews by Marsh, 1987, and by Marsh and Dunkin, 1997) reviewed several large studies of the multivariate relationship between a comprehensive set of background characteristics and SETs. In two such studies (see Marsh, 1987), 16 background characteristics explained about 13% of the variance in the set of SEEQ dimensions, but varied substantially depending on the SEEQ factor. Four background variables could account for most of the explained variance: SETs were correlated with higher prior subject interest, higher expected grades, higher levels of workload/difficulty, and a higher percentage of students taking the course for general interest only. Path analyses demonstrated that prior subject interest had the strongest impact on SETs, and that this variable also accounted for about one-third of the expected-grade effect. Expected grades had a negative effect on workload/difficulty in that students in classes expecting to receive lower grades perceived the course to be more difficult. Even these relatively modest relations, however, need not be interpreted as reflecting bias.

Table 2: Overview of relationships found between student ratings and background characteristics

Background characteristics	Summary of findings
Prior subject interest	Classes with higher interest rate classes more favorably, though it is not always clear if interest existed before start of course or was generated by course/instructor
Expected grade/actual grades	Class-average grades are correlated with class-average SETs, but the interpretation depends on whether higher grades represent grading leniency, superior learning, or pre-existing differences
Reason for taking a course	Elective courses and those with higher percentage taking course for general interest tend to be rated higher
Workload/difficulty	Harder, more difficult courses requiring more effort and time are rated somewhat more favorably
Class size	Mixed findings but most studies show smaller classes rated somewhat more favorably, though some find curvilinear relationships where large classes are also rated favorably
Level of course/year in school	Graduate level courses rated somewhat more favorably; weak, inconsistent findings suggesting upper division courses rated higher than lower division courses
Instructor rank	Mixed findings, but little or no effect
Sex of instructor and/or student	Mixed findings, but little or no effect
Academic discipline	Weak tendency for higher ratings in humanities and lower ratings in sciences, but too few studies to be clear
Purpose of ratings	Somewhat higher ratings if known to be used for tenure/promotion decisions
Administrative conditions	Somewhat higher if ratings not anonymous and instructor present when being completed

Table 2: *(Continued)*

Background characteristics	Summary of findings
Student personality	Mixed findings, but apparently little effect, particularly since different "personality types" may appear in somewhat similar numbers in different classes
Note. For most of these characteristics, particularly the ones that have been more widely studied, some studies have found results opposite to those reported here, while others have found no relationship at all. The size of the relationships often varies considerably, and in some cases even the direction of the relationship, depending upon the particular component of student ratings that is being considered. Few studies have found any of these characteristics to be correlated more than .30 with class-average student ratings, and most reported relationships that were much smaller.	

POTENTIAL BIASES AS A SOURCE OF VALIDITY

Support for a bias hypothesis, as with the study of validity, must be based on a construct validation approach. Indeed, it is ironic that consumers of SET research who have been so appropriately critical of studies claiming to support the validity of SETs have not applied the same level of critical rigor to the interpretation of potential biases in SETs. If a potential biasing factor actually does have a valid influence on teaching effectiveness and this influence is evident in different indicators of teaching effectiveness (e.g., SETs, teacher self-evaluations, student motivation, subsequent course choice, test scores), then it may be possible that the influence reflects support for the validity of SETs (i.e., a valid source of influence in teaching effectiveness is reflected in SETs) rather than a bias. If a potential bias has a substantial effect on specific SET components to which it is most logically related (e.g., class size and individual rapport) but has little or no relation to other SET components (e.g., organization) and this pattern of relations is consistent across multiple methods of measuring teaching effectiveness (e.g., SETs and teacher self-evaluations), again this influence may reflect the validity of SETs rather than a bias. Whereas this still leaves the tricky question of how to control for such differences most appropriately when interpreting SETs, this is a separate question to the most appropriate interpretation of relations between SETs and potential bias

factors. Thus, for example, apparently no one would argue that student learning as articulated in multisection validity studies is a bias to student ratings rather than a source of validity or that student learning should be partialled from SETs to provide a more valid summary of the SETs.

Following Marsh (1987), Centra's (2003) operationalization of bias is consistent with the perspective taken here: "***Bias exists when a student, teacher, or course characteristic affects the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching, such as increased student learning***". Although a thorough discussion of potential biases is beyond the scope of this review (see Marsh, 1984; 1987; Marsh & Dunkin, 1997; Marsh & Roche, 1997; 2000; Marsh, 2001), we briefly present the argument for why many of the most widely posited potential biases to SETs actually support their validity.

Class size. Class size has a small negative relationship with SETs, which is sometimes uncritically interpreted as a bias. However, class size is moderately correlated with factors to which it is most logically related (group interaction and individual rapport, *r*s as large as −0.30). In contrast, it is almost uncorrelated with other SET factors and global ratings and somewhat positively correlated with organization (i.e., teachers are somewhat more organized in large lecture classes than small seminar classes). Importantly, there is a similar pattern of domain specific relations between class size and teacher self-evaluations of their own teaching (Marsh, Overall, & Kesler, 1979a). Also, the class-size effect is nonlinear, such that SETs increase with increasing enrolment beyond an inflection point, such that ratings are as high in very large classes as in small classes. Marsh and Dunkin (1997; also see Marsh, 1987) suggested this reflects more appropriate large-class teaching strategies when class size is very large. Also, students are more likely to enroll in courses taught by the best teachers, suggesting that the direction of causation might be from teaching effectiveness to SETs. Particularly the specificity of the class size effect to SET factors most logically related to this variable, and the similar results for teacher self-evaluations, argues that class size does not bias SETs. Rather, class size has moderate effects on the aspects of effective teaching to which it is most logically related (group interaction and individual rapport) and these effects are accurately reflected in the SETs. Clearly, the nature of class size effect demonstrates that relations must be carefully scrutinized from a construct validity approach before bias interpretations are offered on the basis of correlations.

Prior subject interest. Marsh and Dunkin, 1997; also see Feldman, 1977; Howard & Maxwell, 1980; Howard & Schmeck, 1979) reported that prior subject interest was the most strongly related to SETs of any of the 15 other background variables they considered. In different studies, prior subject interest was consistently more highly correlated with learning/value (*r*s about 0.4) than with any other SEEQ dimensions (rs between 0.3 and −0.12). Instructor self-evaluations of their own teaching were also positively correlated with both their own and their students' perceptions of students' prior subject interest, particularly learning/value. The specificity of the prior subject interest effect to dimensions most logically related to this variable, and the similarity of findings based on SETs and teacher self-evaluations argues that this effect is not a "bias" to SETs. Rather, prior subject interest is a variable that influences some aspects of effective teaching, particularly learning/value, and these effects are accurately reflected in both the SETs and in instructor self-evaluations.

Workload/difficulty. Workload/difficulty is frequently cited by faculty as a potential bias to SETs in the belief that offering less demanding courses will lead to better SETs. However, of critical importance to its interpretation, the direction of the workload/difficulty effect is opposite to that predicted by a bias hypothesis; workload/difficulty is positively—not negatively—correlated with SETS, the direction of the effect generalizing over several different large scale studies based on millions of students, thousands of teachers, and hundreds of universities (see Marsh & Dunkin, 1997; Marsh & Roche, 2000; Marsh, 2001). Overall & Marsh (1979) also reported that instructor self-evaluations of their own teaching effectiveness tended to be positively related to workload/difficulty.

Subsequent research suggests that the workload/difficulty effect is more complicated. For example, Marsh and Roche (2000); Marsh (2001) demonstrated a small non-linear component to the workload effect. For most of the range of the workload/difficulty factor the relation was positive (better SETs associated with higher levels of workload/difficulty). However, they also identified a non-linear component with an inflection point near the top of the workload continuum where SETs levelled off and then decreased slightly. In his recent analysis of 55,549 classes from a diverse sample of universities, Centra (2003) reported a similar nonlinear relation between workload/difficulty and overall teacher evaluations. However, Marsh (2001) found no non-linearity in the positive relation between workload and learning/value. Since the direction of the

workload/difficulty effect was opposite to that predicted as a potential bias, and since this finding is consistent for both SETs and instructor self-evaluations, workload/difficulty does not appear to constitute a bias to SETs.

In a reanalysis of Greenwald and Gillmore's (1997a, 1997b) data, Marsh (2001) found two nearly uncorrelated components of Workload (also see Gillmore & Greenwald, 1994; Frankin & Theall, 1996); good workload was positively related to SETs and learning, but bad workload (time spent that was not valuable) had negative relations. Because the majority of the workload was seen as valuable, the total workload factor was positively related to SETs. Whereas Marsh was able to replicate the non-linear relation between good workload (a positive relation with an inflection point near the top of the workload continuum), the negative relation between SETs and bad workload was linear. Although the results suggest that it is possible to have too much of a good thing, it is important to note that few classes had good workload levels beyond the inflection point. Implications are that most teachers in order to be good teachers – as well as improving their SETs, should increase good workload, but decrease bad workload.

Grading Leniency/Expected Grade Effect

The effect of class-average expected grades and grading leniency on SETs is the most controversial and, perhaps, most misunderstood potential bias in this area of research. Class-average grades are not substantially correlated with SETs. Marsh and Dunkin (1997; Marsh & Roche, 2000) reported that class-average grades correlated .20 with overall teacher ratings in SEEQ research, and this finding is consistent with the extensive review of this relation reported by Feldman (1976a; 1997). Marsh and Dunkin suggested that the best single estimate of the relation between overall teacher rating and expected grades was probably the .2 value reported by Centra and Creech (1976) based on 9,194 class-average responses from a diversity of different universities, courses, settings, and situations. However, Centra (2003), in subsequent research based on a much larger, diverse sample of 55,549 classes, found a slightly lower correlation of only .11. Although the relation is small, it is important to pursue at least three very different interpretations of this relation (Marsh & Dunkin, 1997; Marsh, 2001):

- The ***grading leniency hypothesis*** proposes that instructors who give higher-than-deserved grades will be rewarded with

higher-than-deserved SETs, and this constitutes a serious bias to SETs. According to this hypothesis it is not grades per se that influence SETs, but the leniency with which grades are assigned.

- The ***validity hypothesis*** proposes that better expected grades reflect better student learning, and that a positive correlation between student learning and SETs supports the validity of SETs.
- The ***prior student characteristics hypothesis*** proposes that pre-existing student variables such as prior subject interest may affect student learning, student grades, and teaching effectiveness, so that the expected-grade effect is spurious.

While these and related explanations of the expected-grade effect have quite different implications, actual or expected grades must surely reflect some combination of student learning, the instructor's grading standards, and student characteristics.

In evaluating these alternative interpretations, it is important to emphasize that the critical variable is grading leniency rather than expected grades per se. To the extent that higher expected grades reflect better student learning (instead of lenient grading), the positive relation between class-average expected grades and SETs represents a valid influence, as posited in the validity hypothesis. However, except in special circumstances like the multisection validity study, it is difficult to unconfound the effects of expected grades and grading leniency.

Domain specificity. Marsh and Dunkin (1997; Marsh, 2001; Marsh & Roche, 2000) reported that expected grades correlated between 0 and .30 with different SEEQ factors. The highest correlation is for the learning factor, and this is consistent with the validity hypothesis (that higher grades reflect greater levels of mastery as a result of more effective teaching). Because this relation is reduced substantially by controlling prior subject interest, there is also support for a prior characteristics hypothesis. A similar pattern of results was found with teacher self-evaluations of their own teaching. Expected grades are also moderately correlated with group interaction. This apparently indicates that students tend to receive higher grades in advanced level seminar courses where student-teacher interaction may be better. In support of this interpretation, controlling for class size and class-average year in school substantially reduced this effect, consistent with the prior characteristics hypothesis.

Multisection validity studies. In these studies (reviewed earlier), sections of student in a multi-section course that performed best on a standardized final examination also gave the most favorable SETs.

Because pre-existing differences and grading leniency are largely controlled in these studies, the results provide strong support for the validity hypothesis. Because the size of correlations between actual achievement and SETs in multisection validity studies tends to be as large as or larger than the typical expected-grade correlation, it seems that much of this relation reflects the valid effects of student learning on SETs. This research provides the strongest basis for the interpretation of the expected-grade effect of any research considered here.

Perceived learning. Ideally, it would be useful to control class-average expected grades for the amount students actually learned as an operational definition of grading leniency. However, this is not typically possible in a cross-section of different classes. This is why the results based on multisection validity studies are so important, demonstrating that learning is positively related to SETs when grading leniency (and many other characteristics) are held constant.

In an alternative approach, several research groups (Cashin, 1988; Centra, 1993; Greenwald & Gillmore, 1997a; Howard & Maxwell, 1982) have devised measures of perceived learning as an alternative measure of student learning. These consisted of student self-ratings of progress on specific learning outcomes related to the quality and quantity of learning (e.g., factual knowledge, appreciation, problem solving, real-world application, creativity), rather than teaching effectiveness per se. Consistent with a validity hypothesis—and in direct contradiction to a grading leniency hypothesis—Marsh and Roche (2000) demonstrated that the relation between class-average expected grades and SETs was eliminated once the effect of student perceptions of learning was controlled. Centra (2003) reached a similar conclusion based on his large, diverse sample of 55,549 classes, leading him to conclude that once student ratings of learning outcomes (perceived learning) were controlled, there was no effect of expected grades. Although Marsh and Roche offer cautions about the interpretation of perceived learning as a surrogate of actual student learning, these studies represent one of the few attempts to unconfound expected grades from student learning as must be done if the effects of grading leniency are to be evaluated.

Direct measures of grading leniency. In one of the few studies to measure teacher perceptions of their grading leniency directly, Marsh and Overall (1979) reported that correlations between teacher self-perceptions of their own "grading leniency" (on an "easy/lenient grader" to "hard/strict grader" scale) were significantly correlated with student ratings of grading lenience. Importantly, both student and

teacher ratings of grading leniency were not substantially related to either student and teacher-self evaluations of effective teaching (*r*s between −.16 and .19), except for ratings of workload/difficulty (*r*s of .26 and .28) and teacher self-evaluations of examinations/grading ($r = .32$). In a separate study, Marsh (1976) found that teachers who reported that they were "easy" graders received somewhat (significantly) *lower* overall course and learning/value ratings. Hence, results based on this direct measure of grading leniency argue against the grading leniency hypothesis.

Path analytic approaches. Path analytic studies (see Marsh, 1983, 1987) demonstrate that about one-third of the expected-grade effect is explained in terms of prior subject interest. This supports, in part, the prior characteristics hypothesis.

Experimental field studies. Marsh and Dunkin (1992; Marsh & Roche, 1997; 2000; Marsh, 2001; also see Abrami, Dickens, Perry, & Leventhal, 1980; Centra, 2003; Howard & Maxwell, 1982) reviewed experimental field studies purporting to demonstrate a grading leniency effect on SETs. However, they concluded that this research was flawed in terms of design, grading leniency manipulations, interpretation of the results, and ambiguity produced by deception research. More methodologically adequate studies along the lines of this historical set of studies have not been conducted, because current ethical standards have precluded the type of deception manipulations used in these studies. In contrast, Abrami et al. (1980) conducted what appears to be the most methodologically sound study of experimentally manipulated grading standards in two "Dr. Fox" type experiments (see subsequent discussion) in which students received a grade based on their actual performance but scaled according to different grading standards (i.e., an "average" grade earning a B, C+, or C). Students then viewed a similar lecture, evaluated teacher effectiveness, and were tested again. The grading leniency manipulation had no effect on achievement and weak inconsistent effects on SETs. Whereas the findings do not support a grading-leniency effect, the external validity of the grading manipulation in this laboratory study may also be questioned.

Other approaches. Marsh (1982a) compared differences in expected grades with differences in SETs for pairs of offerings of the same course taught by the same instructor on two different occasions. He reasoned that differences in expected grades in this situation probably represent differences in student performance, since grading standards are likely to remain constant, and differences in prior subject interest were small (for two offerings of the same course) and relatively uncorrelated with

differences in SETs. He found even in this context that students in the more favorably evaluated course tended to have higher expected grades, which argued against the grading leniency hypothesis.

Peterson and Cooper (1980) compared SETs of the same instructors by students who received grades and those who did not. The study was conducted at two colleges where students were free to cross-enrol, but where students from one college were assigned grades but those from the other were not. Whereas class-average grades of those students who received grades were correlated with their class-average evaluations, their class-average evaluations were in substantial agreement with those of students who did not receive grades. Hence, receiving or not receiving grades did not affect SETs. Because grading leniency was unlikely to affect students who did not receive grades, these results suggest that the expected grade effect was not due to grading leniency.

Grade inflation. Even if grading leniency and workload are not significantly related to SETs, a belief that they are may prompt academics to assign grades more leniently and reduce levels of workload, on the assumption that they will be rewarded with higher SETs. In one of the most systematic evaluations of this possibility, Marsh and Roche (2000) evaluated changes in SETs, expected grades, and workload over a 12-year period at one university. Workload did not decrease, but increased slightly over this period; grades neither systematically increased nor decreased over this time period. Although there was a very small increase in SETs over time (0.25% of variance explained), these were not related to changes either in expected grades or workload. However, based on a similar analysis over 40 semesters at a single university, Eiszler (2002) found small increases in both expected grades and SETs, leading him to suggest grade inflation may be related to changes in SETs. Curiously, controlling for cumulative GPA did not substantially reduce the relation between expected grades and SETs, as would be expected if both GPA and expected grades were influenced by grade inflation. Although there were important differences between the two studies (Marsh and Roche based results on class-average means whereas Eiszler, apparently inappropriately, based analyses on semester-average scores aggregated across class-average means), both studies suffered in that they were based on responses from a single university. It would be useful to pursue grading leniency bias in related analyses based upon a large diverse sample of universities such as that used by Centra (1993, 2003) for different purposes.

Summary of grading leniency/expected grades effects. In summary, evidence from a variety of different studies clearly supports the validity and student characteristics hypotheses. Whereas a grading-leniency effect may produce *some* bias in SETs, support for this suggestion is weak, and the size of such an effect is likely to be insubstantial.

THE "DR. FOX" EFFECT

The "Dr. Fox" effect is defined as the overriding influence of instructor expressiveness on SETs, and has been interpreted to mean that an enthusiastic lecturer can "seduce" students into giving favorable evaluations, even though the lecture may be devoid of meaningful content (see Marsh & Dunkin, 1997; Marsh, 1987). In the standard Dr. Fox paradigm, a series of six videotaped lectures—representing three levels of course content (the number of substantive teaching points covered) and two levels of lecture expressiveness (the expressiveness with which a professional actor delivered the lecture)—were all presented by the same actor. Students viewed one of the six lectures, evaluated teaching effectiveness on a multidimensional SET instrument, and completed an achievement test based upon all the teaching points in the high content lecture. In their meta-analysis of this research, Abrami, Leventhal, and Perry (1982) concluded that expressiveness manipulations had substantial impacts on overall SETs and small effects on achievement, whereas content manipulations had substantial effects on achievement and small effects on ratings.

In their reanalysis of the original Dr. Fox studies, Marsh and Ware (1982) identified five SET factors that were differentially affected by the experimental manipulations. Particularly in the condition most like the university classroom, where students were given incentives to do well on the achievement test, the Dr. Fox effect was *not* supported in that: (a) the instructor expressiveness manipulation only affected ratings of instructor enthusiasm, the factor most logically related to that manipulation, and (b) content coverage significantly affected ratings of instructor knowledge and organization/clarity, the factors most logically related to that manipulation. When students were given no added incentives to perform well, instructor expressiveness had more impact on all five student rating factors (though the effect on instructor enthusiasm was still largest), but the expressiveness manipulation also had more impact on student achievement scores than did the content manipulation (i.e., presentation style had more to do with how well students performed on the examination than did the number

of questions that had been covered in the lecture). Hence, as observed in the examination of potential biases to SETs, this reanalysis indicates the importance of considering the multidimensionality of SETs. An effect, which has been interpreted as a "bias" to SETs, seems more appropriately interpreted as support for their validity with respect to one component of effective teaching.

UTILITY OF STUDENT RATINGS

Using a series of related logical arguments, many researchers and practitioners have made the case for why the introduction of a broad institutionally-based, carefully planned program of SETs is likely to lead to the improvement of teaching (see Marsh & Dunkin, 1997; Murray, 1987): (a) SETs provide useful feedback for diagnosing strengths and weaknesses in teaching effectiveness; (b) feedback can provide the impetus for professional development aimed at improving teaching; (c) the use of SETs in personnel decisions provides a tangible incentive to working to improve teaching; and (d) the use of SETs in tenure decisions means that good teachers are more likely to be retained. In support of his argument, Murray (1987; also see Marsh & Dunkin, 1997) summarized results of published surveys from seven universities that asked teachers whether SETs are useful for improving teaching. Across the seven studies, about 80% of the respondents indicated that SETs led to improved teaching. None of these observations, however, empirically demonstrate improvement of teaching effectiveness resulting from SETs.

In most studies of the effects of feedback from SETs, teachers are randomly assigned to experimental (feedback) and one or more control groups; SETs are collected during the course (i.e., midterm ratings); midterm ratings of the teachers in the feedback group are returned to instructors as quickly as possible; and the various groups are compared at the end of the term on a second administration of SETs and sometimes on other variables as well. There are, of course, many variations to this traditional feedback design.

SEEQ Feedback Research

Multisection feedback design. In two early feedback studies with the SEEQ instrument, a multisection feedback design was used in which experimental and control teachers taught different sections of the same multisection course. In the first study, results from an abbreviated form

of the survey were simply returned to faculty; the impact of the feedback was positive, but very modest (Marsh, Fleiner, & Thomas, 1975). In the second study (Overall & Marsh, 1979) researchers actually met with instructors in the feedback group to discuss the evaluations and possible strategies for improvement. In this study, students in the feedback group subsequently performed better on a standardized final examination, rated teaching effectiveness more favorably at the end of the course, and experienced more favorable affective outcomes (i.e., feelings of course mastery, and plans to pursue and apply the subject).

Particularly the Overall and Marsh study was significant, as it was apparently the first to include student learning and other outcomes not easily implemented in studies with diverse courses. However, Hampton and Reiser (2004) replicated the Overall-Marsh multisection design, demonstrating the effectiveness of feedback and consultation compared to a randomly assigned no-feedback control group in terms of instructional practice of the teachers and SETs. Whereas student learning and student motivation were positively correlated with use of instructional activities—a focus of the intervention—differences between experimental and control groups did not reach statistical significance. Even though the multisection feedback design is rarely used, this set of studies highlights important advantages that can be implemented in future research.

Feedback consultation intervention. A critical concern in feedback research is that nearly all of the studies are based on midterm feedback from midterm ratings. This limitation probably weakens effects, in that many instructional characteristics cannot be easily altered within the same semester. Furthermore, Marsh and Overall (1980) demonstrated in their multisection validity study that midterm ratings were less valid than end-of-term ratings.

Marsh and Roche (1993) addressed this issue—as well as others noted in their review of previous research—in an evaluation of a feedback/consultation intervention adapted from Wilson (1986). More specifically, a large, diverse group of teachers completed self-evaluations and were evaluated by students at the middle of Semester 1, and again at the end of Semesters 1 and 2. Three randomly assigned groups received the intervention at midterm of Semester 1, at the end of Semester 1, or received no intervention (control).

A key component of the intervention was a booklet of teaching strategies for each SEEQ factor. Teachers selected the SEEQ factor to be targeted in their individually structured intervention and then selected

the most appropriate strategies from the book of strategies for that SEEQ factor. Ratings for all groups improved over time, but ratings for the intervention groups improved significantly more than those for the control group. The intervention was particularly effective for the initially least effective teachers and the end-of-term feedback was more effective than the midterm feedback.

For the intervention groups (compared to control groups), targeted dimensions improved substantially more than nontargeted dimensions. The study further demonstrated that SET feedback and consultation are an effective means to improve teaching effectiveness and provided a useful procedure for providing feedback/consultation.

Critical features of the Marsh and Roche (1993) intervention were the availability of concrete strategies to facilitate efforts to improve teaching effectiveness in relatively less effective areas that the teacher perceived to be important, the facilitator role adopted by the consultant in this intervention, the personal commitment obtained from the teacher—facilitated by the face-to-face interaction between teacher and consultant, and the multidimensional perspective embodied in feedback booklets and the SEEQ instruments. Fundamental assumptions underlying the logic of the intervention are that teaching effectiveness and SETs are multidimensional, that teachers vary in their effectiveness in different SET areas as well as in perceptions of the relative importance of the different areas, and that feedback specific to particular SET dimensions is more useful than feedback based on overall or total ratings, or that provided by SET instruments which do not embody this multidimensional perspective. Indeed, this intervention can only be conducted with a well-designed, multidimensional instrument like SEEQ and feedback booklets specifically targeted to the SEEQ factors.

Meta-analyses of Feedback Research

In his classic meta-analysis, Cohen (1980) found that instructors who received midterm feedback were subsequently rated about one-third of a standard deviation higher than controls on the total rating (an overall rating item or the average of multiple items), and even larger differences were observed for ratings of instructor skill, attitude toward subject, and feedback to students. Studies that augmented feedback with consultation produced substantially larger differences, but other methodological variations had little effect (also see L'Hommedieu,

Menges, & Brinko, 1990). The most robust finding from the feedback research reviewed here is that consultation augments the effects of written summaries of SETs, but insufficient attention has been given to determine the type of consultative feedback that is most effective.

L'Hommediu, Menges, and Brinko (1990) critically evaluated feedback studies. They concluded that the overall effect size attributable to feedback was probably attenuated due to a number of characteristics of the traditional feedback paradigm, and developed methodological recommendations for future research. Among their many recommendations, they emphasized the need to: use a larger number of instructors; more critically evaluate findings within a construct validity framework, as emphasized by Marsh (1987); more critically evaluate the assumed generalizability of midterm feedback to end-of-term feedback; base results on well-standardized instruments such as SEEQ; and use more appropriate no-treatment controls. In their meta-analysis, they considered three forms of feedback that differed systematically in their effect sizes: written feedback consisting of printed summaries of SETs (Mean effect = .18); personal feedback consisting of summary material delivered in person, sometimes accompanied by interpretations, discussion, and advice (mean effect = .25); and consultative feedback that combines SET feedback and professional development (mean effect = .87). Consistently with Cohen (1980) they concluded that "the literature reveals a persistently positive, albeit small, effect from written feedback alone and a considerably increased effect when written feedback is augmented with personal consultation" (1990, p. 240), but that improved research incorporating their suggestions would probably lead to larger, more robust effects.

More recently, Penny and Coe (2004) conducted a meta-analysis of 11 studies that specifically contrasted consultative feedback based on a dialogue with a consultant, with randomly assigned control groups. They found an overall effect size of .69, consistent with earlier results. Although they did not find significant study-to-study variation, they pursued a systematic evaluation of moderator effects. The largest effects were associated with the use of a well-standardized rating instrument, and consultations that incorporated a consultative or educational approach (rather than a purely diagnostic approach that focused on interpretation of the ratings). Whereas they offered heuristic recommendations about providing consultation, their sample size was so small that highlighted differences rarely achieved statistical significance. As advocated by Penny and Coe, there is need for further research to explore more fully their recommendations.

Other Uses of SETs

Personnel decisions. In research reviewed by Marsh and Dunkin (1997) there is clear evidence that the importance and usefulness of SETs as a measure of teaching effectiveness have increased dramatically during the last 60 years. Despite the strong reservations of some, faculty are apparently in favor of the use of SETs in personnel decisions—at least in comparison with other indicators of teaching effectiveness. In order to evaluate experimentally the importance of teaching effectiveness in personnel decisions, Leventhal, Perry, Abrami, Turcotte and Kane (1981), and Salthouse, McKeachie, and Lin (1978) composed fictitious summaries of faculty performance that systematically varied reports of teaching and research effectiveness, and also varied the type of information given about teaching (chairperson's report, or chairperson's report supplemented by summaries of SETs). Both studies found reports of research effectiveness to be more important in evaluating total faculty performance at research universities, although Leventhal et al. found teaching and research to be of similar importance across a broader range of institutions. While teaching effectiveness as assessed by the chairperson's reports did make a significant difference in ratings of overall faculty performance, neither study found that supplementing the chairperson's report with SETs made any significant difference. However, neither study considered SETs alone, or even suggested that the two sources of evidence about teaching effectiveness were independent. Information from the ratings and the chairperson's report was always consistent, so that one was redundant, and it would be reasonable for subjects in these studies to assume that the chairperson's report was at least partially based upon SETs. These studies demonstrate the importance of reports of teaching effectiveness, but apparently do not test the impact of SETs.

In other research related to the use of SETs for personnel decisions, Franklin and Theall (1989) argue that SETs can be misused or misinterpreted when making personnel decisions. This introduces another source of invalidity in the interpretation of SETs—even if the SETs are reliable and valid in relation to the traditional psychometric criteria considered in this chapter. Here, as in other areas of research on how SETs are most appropriately used to enhance their utility, there is a dearth of relevant research.

Usefulness in Student Course Selection. Little empirical research has been conducted on the use of ratings by prospective students in the selection of courses. UCLA students reported that the Professor/Course

Evaluation Survey was the second most frequently read of the many student publications, following the daily, campus newspaper (Marsh, 1987). Similarly, about half the Indiana University students in Jacob's (1987) study generally consulted published ratings prior to taking a course. Leventhal, Abrami, Perry and Breen (1975) found that students say that information about teaching effectiveness influences their course selection. Students who select a class on the basis of information about teaching effectiveness are more satisfied with the quality of teaching than are students who indicate other reasons (Centra & Creech, 1976; Leventhal, Abrami, & Perry, 1976; also see Babad et al. 1999; Perry et al., 1979). In an experimental field study, Coleman and McKeachie (1981) presented summaries of ratings of four comparable political science courses to randomly selected groups of students during preregistration meetings. One of the courses had received substantially higher ratings, and it was chosen more frequently by students in the experimental group than by those in the control group. Hence, apparently SETs are useful for students in the selection of instructors and courses.

USE OF NORMATIVE COMPARISONS

In many programs, the SET raw scores are compared with those obtained by large representative groups of classes in order to enhance the usefulness of the feedback. Although arguments for and against the use of normative comparisons and related issues have tended to be overly simplistic, this is a complicated issue fraught with theoretical, philosophical, and methodological quagmires for the unsuspecting. Here I distinguish between three related issues: use of norms to enhance the usefulness of SETs, the construction of norms to control potential biases to SETs, and the setting of standards.

Enhancing the usefulness of SETs. Traditionally, one of the key differences between broad, institutionally developed programs of SETs and ad hoc instruments has been the provision of normative comparisons. Marsh (1987; Marsh & Dunkin, 1997), like many others, argued that the usefulness of the raw scores is enhanced by appropriate normative comparisons, because raw score ratings on SET factors, global rating items, and specific rating items are likely to be idiosyncratic to the particular wording of the item. Furthermore, scores on different items and SET factors are not directly comparable in the original raw score metric. The metric underlying raw scores is not well defined and varies from item to item (and from factor to factor).

Hence, the normative comparisons provide information on how ratings on different SET factors for a given teacher compare with those based on a suitably constructed normative group of teachers and classes, and how scores from different SET items and factors for the same teacher compare to each other.

McKeachie (1996) provoked an interesting debate about the desirability of normative comparisons. Although he did not necessarily question the potential usefulness of appropriate normative comparisons, he argued that the unintended negative consequences might outweigh potential benefits. Thus, because nearly all class-average student ratings fall above the mid-point of the rating scale (e.g., above 3.0 on a typical 1–5 scale in which 5 is the highest rating), teachers can feel good about themselves even if they fall below the normative average response. According to McKeachie, if teachers are demoralized by low ratings, then the consequences may be more negative than if this supplemental information were not made available.

My perspective, although sympathetic with the potential dangers of social comparison on self-perceptions and implications for future performance (Marsh & Hau, 2003) is quite different. Indeed, I argue that it may be unethical—certainly patronizing—to deny teachers potentially useful information based on the assumption that we know what is best for them. Gillmore (1998), also arguing for the usefulness of normative comparisons, suggested that a strategic compromise might be to provide extensive norms via the web that are readily accessible, but not to provide these normative comparisons as part of the standard feedback presented to academics. My recommendation is that raw scores and scores normed in relation to at least one appropriately constructed normative comparison group should be included as part of the feedback given to teachers (Marsh, 1987).

Control for potential biases. Even if the usefulness of normative comparisons is accepted, there are critical issues involved in the construction of appropriate norms. For example, some researchers advocate that SETs should be adjusted for potential biases to the SETs (e.g., class size, expected grades, prior subject interest, workload/difficulty) based on multiple regression. I also dispute the appropriateness of this approach on methodological and philosophical grounds. As I have argued here, bias can only be inferred in relation to a well-defined operational definition of bias. At least based on the definition of bias used here (also see Centra, 2003), there is little support for any of these characteristics as biases to SETs. The adjustment rationale may, perhaps, be more appropriate in a relation

to a definition of bias based on a fairness notion. Hence, to the extent that some characteristic is not under the control of the teacher, it might be "fair" to adjust for this characteristic. Logically, such adjustments should be based on characteristics that are readily discernible prior to the start of actual instruction, to avoid potential confounding of factors that influence the SETs, rather than being influenced by teaching effectiveness. This would preclude, for example, adjustments to class-average actual or expected grades, which are clearly under the control of the teacher and influenced by teaching effectiveness. Whereas it may, for example, be reasonable to adjust for prior subject interest, it could be argued that some of the class-average prior subject interest ratings collected at the end of the course might reflect effective teaching in addition to the effect of prior subject interest in ratings of this construct collected prior to the start of the class. Even a characteristic such as class size is not completely unproblematic if students choose teachers on the basis of teaching effectiveness, such that teaching effectiveness causes class size.

An alternative, somewhat more acceptable compromise is to construct separate normative comparison groups of similar courses. Thus, for example, Marsh (1987) described how SEEQ ratings are normed in relation to courses from three groups (Teaching Assistants, undergraduate courses taught by regular teachers, and graduate level courses) and there is provision—subject to adequate sample sizes—to form norm groups specific to a particular discipline (Marsh & Roche, 1994). This solution, although overcoming some of the problems associated with statistical adjustment, would still be problematic if norm groups were formed on the basis of class characteristics that reflect teaching effectiveness instead of (or in addition to) a source of bias or unfairness. Thus, for example, I would argue against the construction of norm groups based on class-average expected grades.

Other standards of comparison. Particularly when normative comparisons are presented, there is an emphasis on how the ratings of a teacher compare with those obtained by other teachers. This social comparison emphasis, as noted by McKeachie (1996), might have unintended negative consequences. In contrast, rating profiles (see earlier discussion of profile analyses) focus more specifically on the relative strengths and weaknesses in relation to the different SEEQ factors. Whereas the "level" of the profile for any given factor reflects a normative comparison with an appropriate norm group, the differences between the different factors (the "shape" component in profile analyses) are a more salient feature of this graphical presentation.

So long as SET programs retain appropriate archives over an extended period of time, it is also possible to use the graphical profile to compare one set of ratings with those based on previous ratings by the same teacher. Thus, for example, the profile graphs presented by Marsh and Bailey (1993) were based on many sets of ratings by the same teacher. They noted, however, it would be easy to extend these graphs to show the current set of ratings simultaneously, to allow the teacher to easily evaluate progress in relation to his or her own previous performance—further de-emphasizing the normative comparisons with other teachers.

Focusing on the use of different SET factors as a basis of improvement, Marsh and Roche (1994) asked teachers to focus improvement efforts on a specific SET factor—one on which their ratings were low relative to other SEEQ factors in a multidimensional profile based on their previous ratings, and one that they rated as important in self-evaluations of their own teaching.

In summary, alternative frames of reference against which to judge SETs include the performance of other teachers, previous ratings by the same teacher, or the ratings on one SET factor in relation to those of other SET factors. Particularly when the focus of the SET program is on the improvement of teaching effectiveness, it is appropriate for teachers to set their own standards for what they hope to accomplish in relation to ratings of other teachers, their own previous ratings, or even the relative performance on different SET factors.

Goals and standards of comparison. In considering the use of norms, it is important to distinguish between normative comparisons and standards of what is acceptable, appropriate, or good benchmarks of effective teaching. For present purposes we focus on the use of SETs but it is important to emphasize that there are many criteria of effective teaching—some of which are idiosyncratic to a particular course. A critical aspect of feedback relates to the goals or intended standards of performance. Effective goals involve challenge and commitment (Hattie, 2003; Hattie, Biggs & Purdie, 1996). They inform individuals "as to what type or level of performance is to be attained so that they can direct and evaluate their actions and efforts accordingly. Feedback allows them to set reasonable goals and to track their performance in relation to their goals so that adjustments in effort, direction, and even strategy can be made as needed" (Locke & Latham, 1990, p. 23). As a consequence of feedback, it is critical for teachers to set appropriately challenging goals. When goals have appropriate challenge and teachers are committed to these goals, then a clearer understanding of

the appropriate success criteria is likely to be understood and shared. This focus on having teachers select the most appropriate areas to improve teaching, using prior SETs as a basis of comparison for evaluating improvement, fostering a sense of commitment in achieving improved teaching effectiveness in relation to specific targeted factors, and providing concrete strategies on how to achieve this goal is at least implicit in SET feedback studies (e.g., Marsh & Roche, 1994). However, there is clearly a need to integrate more fully lessons on effective forms of goal setting and feedback (e.g., Hattie, 2003; Hattie et al., 1996; Locke & Latham, 1990) into SET research.

Summary. In summary, normative comparisons provide a valuable additional source of information in the interpretation of SETs. Rather than denying teachers this valuable source of information, it is more appropriate to develop normative comparisons that are more useful to teachers. Here, for example, I emphasize the usefulness of multidimensional profiles that focus on a comparison of relative strengths and weakness for the different components of teaching effectiveness, and on longitudinal comparisons that focus on changes over time in the ratings of the same teacher. Nevertheless, the theoretical, methodological, and philosophical issues inherent in the construction of appropriate normative comparisons are important areas in need of further research. Clearly the appropriate construction of normative comparison groups is an important issue that has received surprisingly little research. Hence, instead of getting rid of norms, we need to enhance their usefulness.

SUMMARY OF STUDIES OF THE UTILITY OF STUDENT RATINGS

With the possible exception of feedback studies on improving teaching based on midterm ratings, studies of the usefulness of SETs are infrequent and often anecdotal. This is unfortunate, because this is an area of research that can have an important and constructive impact on policy and practice. Critical, unresolved issues in need of further research were identified.

- For administrative decisions, SETs can be summarized by responses to a single global rating item, by a single score representing an optimally-weighted average of specific components, or a profile of multiple components, but there is limited research on which is most effective.

- Debates about whether SETs have too much or too little impact on administrative decisions are seldom based upon any systematic evidence about the amount of impact they actually do have.
- Researchers often indicate that SETs are used as one basis for personnel decisions, but there is a dearth of research on the policy practices that are actually employed in the use of SETs.
- Rather than to deny the usefulness of normative comparisons, more research is needed on the most appropriate strategies to construct normative comparisons that enhance the usefulness of SETs. Whereas normative comparisons are an important basis of comparison, too little work has been done on alternative standards of effective teaching.
- A plethora of policy questions exists (e.g., how to select courses to be evaluated, the manner in which rating instruments are administered, who is to be given access to the results, how ratings from different courses are considered, whether special circumstances exist where ratings for a particular course can be excluded, either a priori or post-hoc, whether faculty have the right to offer their own interpretation of ratings, etc.), which are largely unexplored despite the wide use of SETs.
- Anecdotal reports often suggest that faculty find SETs useful, but there has been little systematic attempt to determine what form of feedback to faculty is most useful (although feedback studies do support the use of services by an external consultant), and how faculty actually use the results which they do receive.
- Some researchers have cited anecdotal evidence for negative effects of SETs (e.g., lowering grading standards or making courses easier) but these are also rarely documented in systematic research. Critics suggest that SETs lead to more conservative teaching styles, but Murray (1987) counters that highly rated teachers often use nontraditional approaches and that teaching is less traditional today than it was before SETs were used widely.
- McKeachie (personal communication, 19 March, 1991) noted that SETs are typically used constructively, encouraging instructors to think of alternative approaches and to try them out. He also suggested, however, that if SETs are used destructively so that teachers feel that they are in competition with each other—"that they must always be wary of the sword of student ratings hanging over their head"—poor ratings may increase

anxiety and negative feelings about students so that teaching and learning may suffer. Again, research is needed to examine whether teachers react constructively or destructively to SETs and whether there are individual differences that influence these reactions.

- Although SETs are sometimes used by students in their selection of courses, there is little guidance about the type of information which students want and whether this is the same as is needed for other uses of SETs. Typically, publication of SET results is a highly controversial issue.

These, and a wide range of related questions about how SETs are actually used and how their usefulness can be enhanced, provide a rich field for further research.

USE OF SETS TO BENCHMARK UNIVERSITIES: QUALITY ASSURANCE

In Australia, UK, Hong Kong, and many other countries, there are major governmental initiatives to enhance the accountability of universities by collecting comparable data for purposes of benchmarking and comparing different universities, different disciplines, and different disciplines within universities. Thus, for example, highly standardized and audited measures of research productivity are sometimes used to rank universities and disciplines within universities that determine, in part, the research funding that different universities receive. Hence, the Australian government commissioned the development and evaluation of the Postgraduate Research Experience Questionnaire (PREQ) to provide a multidimensional measure of the experience of postgraduate research students. An initial trial of the PREQ led to very positive recommendations about its psychometric properties (factor structure and reliability) and its potential usefulness as part of a large-scale national benchmarking exercise for Australian universities (Marsh et al., 2002). However, the unit of analysis was a critical issue in this research, as the intended focus was on the overall postgraduate experience at the broad level of the university, and disciplines within a university, rather than the effectiveness of individual supervisors. Indeed, students were specifically asked not to name their supervisor, and some of the factors focused on departmental or university level issues.

Marsh, Rowe, and Martin (2002) evaluated PREQ, a multidimensional measure of PhD and research Masters students' evaluation of the quality of research supervision, that was administered to graduates ($n = 1832$) from 32 Australian and New Zealand Universities. At the level of the individual student, responses had reasonable psychometric properties (factor structure and internal consistency estimates of reliability). Consistent with the potential use of these instruments to benchmark the quality of supervision across all Australian universities, Marsh et al. evaluated the extent to which responses reliably differentiated between universities, academic disciplines, and disciplines within universities. Based on fitting two-level (individual student, university) and three-level (individual student, discipline, university) multilevel models, the responses failed to differentiate among universities, or among disciplines within universities. Although there were small differences between ratings in a few disciplines, even these small differences were consistent across different universities. The results demonstrate that PREQ responses that are adequately reliable at one level (individual student) may have little or no reliability at another level (university). Marsh et al. concluded that PREQ responses should not be used to benchmark Australian universities or disciplines within universities. Furthermore, Marsh, et al. argued that PREQ responses, as presently formulated, were unlikely to be useful for most other conceivable purposes.

The most salient finding of this study was that PREQ ratings did not vary systematically between universities, or between disciplines within universities. This has critically important methodological and substantive implications for the potential usefulness of the PREQ ratings. Because there was no significant variation at the university level, it follows that the PREQ ratings were completely unreliable for distinguishing between universities. This clearly demonstrates why it is important to evaluate the reliability of responses to a survey instrument in relation to a particular application and the level of analysis that is appropriate to this application. Although PREQ ratings were reliable at the level of individual students, these results are not particularly relevant for the likely application of the PREQ ratings to discriminate between universities. Whereas SET research suggests that PREQ ratings might be reliable at the level of the individual supervisor, the number of graduating PhD students associated with a given supervisor in any one year might be too small to achieve acceptable levels of reliability, and there are important issues of anonymity and confidentiality. There are apparently no comparable studies of the ability of SET ratings to

differentiate between universities or even departments within universities, but I suspect that the results would be similar.

Substantively, the Marsh, Rowe, and Martin (2002) study questions is the potential usefulness of PREQ ratings in benchmarking different universities, although the Australian government is continuing to use them for this purpose. More generally, it calls into question research or practice that seeks to use SETs as a basis for comparing universities as part of a quality assurance exercise. Clearly this is an area in need of further research. Although the existence of an effective SET program coupled with a program to improve teaching effectiveness is clearly a relevant criteria upon which to evaluate a university in relation to quality assurance, it is not appropriate – or at least premature – to use SETs from different universities to evaluate differences in teaching effectiveness at those universities.

HOW SETs SHOULD NOT BE USED

There is broad acceptance that SETs should not be the only measure of teaching effectiveness used, particularly for personnel decisions. Indeed, there are a number of areas in which results based on SETs should be supplemented with other sources of information. Thus, for example, whereas students provide relevant information about the currency of materials and the breadth of content coverage, this is clearly an area in which peer evaluations of the course syllabus and reading list should provide major input.

There are other areas where SETs, perhaps, should not be used at all. Particularly for universities with a clear research mission, a major component of the personnel decisions should be based on appropriate indicators of research. The results of the present investigation indicate that SETs—particularly at the level of the individual teacher—are nearly unrelated to research productivity. Highly productive researchers are equally likely to be good teachers as poor teachers. Hence, SETs should not be used to infer research productivity. However, because most universities have at least an implicit mission to enhance the nexus between teaching and research, this is an appropriate area in which to seek student input, and warrants further research.

At least for the type of items used on instruments like SEEQ and dimensions like those summarized by Feldman (1987; also see Table 1), SETs reflect primarily the teacher who does the teaching rather than the particular course that is taught. Even when students are specifically asked to evaluate the course rather than the teacher (i.e., overall course

ratings as opposed to overall instructor ratings) the ratings are primarily a function of the teacher and do not vary systematically with the course. These results greatly enhance the usefulness of SETs for purposes of the evaluation of teachers, but seriously undermine their usefulness for purposes of the evaluation of courses independent of the teacher. It may be possible to construct different items reflecting different dimensions that are useful for evaluations of courses rather than the teacher, and there may be idiosyncratic circumstances in which differences between courses are much more important than particular teachers, but the SET research does not appear to provide support for these suppositions. On this basis, I recommend that SETs not be used to evaluate courses independently of the teachers who teach the course.

Increasingly, SETs are being incorporated into quality assurance exercises like that based on the PREQ research. Clearly, it is appropriate to evaluate the quality of the SET program instituted by a university and provision for systematic programs to improve teaching effectiveness. A useful contribution would be to develop appropriate checklists for indicators of an effective SET program for use in quality assurance exercises. However, the PREQ research suggests that it would be inappropriate to use SETs to evaluate the quality of teaching across different universities or even departments within universities. Nevertheless, recommendations based on ratings of research supervision by PhD students are not a fully satisfactory basis of inference about SETs based on classroom teaching. Particularly given the exciting advances in the application of multilevel modeling, there are likely to be new developments in this area. However, pending results of new research, I recommend that the actual numerical ratings based on SETs should not be used to compare universities in quality assurance exercises.

OVERVIEW, SUMMARY AND IMPLICATIONS

Research described in this chapter demonstrates that SETs are multidimensional, reliable and stable, primarily a function of the instructor who teaches a course rather than the course that is taught, relatively valid against a variety of indicators of effective teaching, relatively unaffected by a variety of potential biases, and seen to be useful by faculty, students, and administrators. I recommend that researchers adopt a construct validation approach in which it is recognised that: effective teaching and SETs designed to reflect teaching effectiveness are multidimensional; no single criterion of effective teaching is sufficient; and tentative interpretations of relations with validity criteria and

with potential biases should be evaluated critically in different contexts and in relation to multiple criteria of effective teaching. In contrast to SETs, however, there are few other indicators of teaching effectiveness whose use is systematically supported by research findings. As noted by Cashin (1988), "student ratings tend to be statistically reliable, valid, and relatively free from bias, probably more so than any other data used for faculty evaluation" (p. 5). Of particular importance, the review demonstrates that the combined use of a good evaluation instrument like SEEQ and an effective consultation procedure like that adapted from Wilson (1986) can lead to improved university teaching.

Despite the many positive features identified in this review, there are a host of critical, unanswered questions in need of further research. Particularly discouraging is the observation that—with a few major exceptions—SET research during the last decade seems not to have adequately addressed these issues that were clearly identified a decade ago. Indeed, relative to the heydays of SET research in the 1980s, the amount and quality of SET research seems to have declined. This is remarkable, given the ongoing controversies that SETs continue to incite, the frequency of their use in universities in North America and, increasingly, throughout the world, and important advances in statistical and methodological tools for evaluating SETs.

Particularly critical issues have to do with the appropriate form to present SETs to enhance their usefulness for formative and summative feedback, and how most appropriately to integrate SETs into programs to enhance teaching effectiveness. Perhaps the most damning observation is that most of the emphasis on the use of SETs is for personnel decisions rather than on improving teaching effectiveness. Even here, however, good research on how SETs are most appropriately used to inform personnel decisions is needed. Although much work is needed on how best to improve teaching effectiveness, it is clear that relatively inexpensive, unobtrusive interventions based on SETs can make a substantial difference in teaching effectiveness. This is not surprising, given that university teachers typically are given little or no specialized training on how to be good teachers and apparently do not know how to fully utilize SET feedback without outside assistance. Why do universities continue to collect and disseminate potentially demoralising feedback to academics without more fully implementing programs to improve teaching effectiveness? Why is there not more SET research on how to enhance the usefulness of SETs as part of a program to improve university teaching? Why have there been so

few intervention studies in the last decade that address the problems identified in reviews of this research conducted a decade ago?

Indeed, it is remarkable that after nearly a century of extensive research, there is apparently no general theory of college teaching that has arisen from SET research. Clearly, the science to support a theory of college teaching does exist in the communal agreement on the key dimensions of effective teaching, appropriate outcome variables, well-established research paradigms, design features, statistical analyses, meta-analyses, and the accumulated findings from a diverse range of laboratory, quasi-experimental, field, longitudinal, and correlational studies. Given the ongoing interest in the science, analysis, interpretation and uses of SETs the time for this type of unified theory building is long overdue.

REFERENCES

Abrami, P.C., and d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness: Generalizability of $N = 1$ research, Comment on Marsh (1991). *Journal of Educational Psychology* 30: 221–227.

Abrami, P.C., d'Apollonia, S., and Cohen, P.A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology* 82: 219–231.

Abrami, P.C., d'Apollonia, S., and Rosenfield, S. (1997). The dimensionality of student ratings of instruction: What we know and what we do not. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research* (Vol. 11, pp. 213–264). New York: Agathon.

Abrami, P.C., d'Apollonia, S., and Rosenfield, S. (March, 1993). *The Dimensionality of Student Ratings of Instruction*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

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

Abrami, P.C., Dickens, W.J., Perry, R.P., and Leventhal, L. (1980). Do teacher standards for assigning grades affect student evaluations of instruction? *Journal of Educational Psychology* 72: 107–118.

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

Apodaca, P., and Grad, H. (2005). The dimensionality of student ratings of teaching: integration of uni- and multidimensional models. *Studies in Higher Education* 30: 723–748.

Babad, E., Darley, J., and Kaplowitz, H. (1999). Developmental aspects in students' course selection. *Journal of Educational Psychology* 91: 157–168.

Blackburn, R.T. (1974). The meaning of work in academia. In J.I. Doi (ed.), Assessing faculty effort. New Directions for Institutional Research (Vol. 2, pp. 75–99). San Francisco: Jossey-Bass.

Braskamp, L.A., Brandenburg, D.C., and Ory, J.C. (1985). *Evaluating Teaching Effectiveness: A Practical Guide*. Beverly Hills, CA: Sage.

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

Braskamp, L.A., and Ory, J.C. (1994). *Assessing Faculty Work: Enhancing Individual and Institutional Performance*. San Francisco, Jossey-Bass.

Cadwell, J., and Jenkins, J. (1985). Effects of the semantic similarity of items on student ratings of instructors. *Journal of Educational Psychology* 77: 383–393.

Cashin, W.E. (1988). *Student Ratings of Teaching. A Summary of Research*. (IDEA paper No. 20). Kansas State University, Division of Continuing Education. (ERIC Document Reproduction Service No. ED 302 567).

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

Centra, J.A. (1979). *Determining Faculty Effectiveness*. San Francisco, CA: Jossey-Bass.

Centra, J.A. (1983). Research productivity and teaching effectiveness. *Research in Higher Education* 18: 379–389.

Centra, J.A. (1989). Faculty evaluation and faculty development in higher education. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research. Supplementary* (Vol. 5,pp. 155–179). New York: Agathon Press.

Centra, J.A. (1993). *Reflective Faculty Evaluation*. San Francisco, CA: Jossey-Bass.

Centra, J.A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education* 44(5): 495–518.

Centra, J.A., and Creech, F.R. (1976). *The Relationship between Student, Teacher, and Course Characteristics and Student Ratings of Teacher Effectiveness* (Project Report 76–1). Princeton, NJ: Educational Testing Service.

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

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

Cohen, P.A. (April, 1987). *A Critical Analysis and Reanalysis of the Multisection Validity Meta-analysis*. Paper presented at the 1987 Annual Meeting of the American Educational Research Association, Washington, DC (ERIC Document Reproduction Service No. ED 283 876).

Coleman, J., and McKeachie, W.J. (1981). Effects of instructor/course evaluations on student course selection. *Journal of Educational Psychology* 73: 224–226.

Costin, F., Greenough, W.T., and Menges, R.J. (1971). Student ratings of college teaching: Reliability, validity and usefulness. *Review of Educational Research* 41: 511–536.

Cranton, P.A., and Hillgartner, W. (1981). The relationships between student ratings and instructor behavior: Implications for improving teaching. *Canadian Journal of Higher Education* 11: 73–81.

Cranton, P., and Smith, R.A. (1990). Reconsidering the unit of analysis: A model of student ratings of instruction. *Journal of Educational Psychology* 82: 207–212.

Cronbach, L.J. (1958). Proposals leading to analytic treatment of social perception scores. In R. Tagiuri and L. Petrullo (eds.), *Person Perception and Interpersonal Behavior* (pp. 351–379). Stanford University Press.

de Wolf, W.A. (1974). *Student Ratings of Instruction in Post Secondary Institutions: A Comprehensive Annotated Bibliography of Research Reported Since 1968* (Vol. 1). University of Washington Educational Assessment Center. Educational Assessment Center.

Doyle, K.O. (1975). *Student Evaluation of Instruction*. Lexington, MA: D. C. Heath.

Doyle, K.O. (1983). *Evaluating Teaching*. Lexington, MA: Lexington Books.

Eiszler, C.F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education* 43(4): 483–501.

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

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

Feldman, K.A. (1977). Consistency and variability among college students in rating their teachers and courses. *Research in Higher Education* 6: 223–274.

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

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

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

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

Feldman, K.A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education* 24: 139–213.

Feldman, K.A. (1987). Research productivity and scholarly accomplishment: A review and exploration. *Research in Higher Education* 26: 227–298.

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

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

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

Feldman, K.A. (1990). An afterword for "the association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies". *Research in Higher Education* 31: 315–318.

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

Feldman, K.A. (1993). College Students' Views of Male and Female College Teachers. Part II-Evidence from Students' Evaluations of Their Classroom Teachers. *Research in Higher Education* 34: 151–211.

Feldman, K.A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R.P. Perry and J.C. Smart, (eds.), *Effective Teaching in Higher Education: Research and Practice* (pp. 368–395). New York: Agathon.

Feldman, K.A. (1998). Reflections on the effective study of college teaching and student ratings: one continuing quest and two unresolved issues. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research* (pp. 35–74). New York: Agathon Press.

Franklin, J.L., and Theall, M. (1989). *Who Reads Ratings. Knowledge, Attitudes, and Practices of Users of Student Ratings of Instruction*. Paper presented at the 70th annual meeting of the American Educational Research Association. San Francisco: March 31.

Franklin, J., and Theall, M. (1996). *Disciplinary Differences in Sources of Systematic Variation in Student Ratings of Instructor Effectiveness and Students' Perceptions of the Value of Class Preparation Time: A Comparison of Two Universities' Ratings Data*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Gilmore, G.M. (1988). *Grades, Ratings and Adjustments*. Instructional Evaluation and Faculty Development (available on internet: http://www.umanitoba.ca/uts/sigfted/backissues.php, 25 August, 2006).

Gillmore, G.M., and Greenwald, A.G. (1994). *The Effects of Course Demands and Grading Leniency on Student Ratings of Instruction*. Office of Educational Assessment (94–4), University of Washington, Seattle.

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

Greenwald, A.G., and Gillmore, G.M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist* 52: 1209–1217.

Greenwald, A.G., and Gillmore, G.M. (1997b). No Pain, No Gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology* 89: 743–751.

Hampton, S.E., and Reiser, R.A. (2004). Effects of a theory-based feedback and consultation process on instruction and learning in college classrooms. *Research in Higher Education* 45(5): 497–527.

Harrison, P.D., Douglas, D.K., and Burdsal, C.A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education* 45(3): 311–323.

Harrison, P.D., More, P.S., and Ryan, J.M. (1996) College student's self-insight and common implicit theories in ratings of teaching effectiveness. *Journal of Educational Psychology* 88: 775–782.

Hattie, J.A. (2003). Why is it so difficult to enhance self-concept in the classroom: The power of feedback in the self-concept–achievement relationship. Paper presented at the International SELF conference, Sydney, Australia.

Hattie, J.A., Biggs, J., and Purdie, N. (1996). Effects of learning skills intervention on student learning: A meta-analysis. *Review of Research in Education* 66: 99–136.

Hattie, J., and Marsh, H.W. (1996). The relationship between research and teaching—a meta-analysis. *Review of Educational Research* 66: 507–542.

Hativa, N. (1996). University instructors' ratings profiles: Stability over time, and disciplinary differences *Research In Higher Education* 37: 341–365.

Hobson, S.M., and Talbot, D.M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching* 49(1): 26–31.

Howard, G.S., Conway, C.G., and Maxwell, S.E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology* 77: 187–196.

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

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

Howard, G.S., and Schmeck, R.R. (1979). Relationship of changes in student motivation to student evaluations of instruction. *Research in Higher Education 10*: 305–315.

Howell, A.J., and Symbaluk, D.G. (2001). Published student ratings of instruction: Revealing and reconciling the views of students and faculty. *Journal of Educational Psychology* 93: 790–796.

Jackson, D.L., Teal, C.R., Raines, S.J., Nansel, T.R., Force, R.C., and Burdsal, C.A. (1999). The dimensions of students' perceptions of teaching effectiveness. *Educational and Psychological Measurement* 59: 580–596.

Jacobs, L.C. (1987). *University Faculty and Students' Opinions of Student Ratings*. Bloomington, IN: Bureau of Evaluative Studies and Testing. (ERIC Document Reproduction Service No. ED 291 291).

Kane, M.T., Gillmore, G.M., and Crooks. T.J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement* 13: 171–184.

Kember, D., Leung, D.Y.P., and Kwan, K.P. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education* 27: 411–425.

Kulik, J.A., and McKeachie, W.J. (1975). The evaluation of teachers in higher education. *Review of Research in Higher Education* 3: 210–240.

L'Hommedieu, R., Menges, R.J., and Brinko, K.T. (1990). Methodological explanations for the modest effects of feedback. *Journal of Educational Psychology* 82: 232–241.

Leventhal, L., Abrami, P.C., and Perry, R.P. (1976). Teacher rating forms: Do students interested in quality instruction rate teachers differently? *Journal of Educational Psychology* 68: 441–445.

Leventhal, L., Abrami, P.C., Perry, R.P., and Breen L.J. (1975). Section selection in multi-section courses: Implications for the validation and use of student rating forms. *Educational and Psychological Measurement* 35: 885–895.

Leventhal, L., Perry, R.P., Abrami, P.C., Turcotte, S.J.C., and Kane, B. (1981, April). *Experimental Investigation of Tenure/Promotion in American and Canadian Universities*. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles.

Lin, Y., McKeachie, W.J., and Tucker, D.G. (1984). The use of student ratings in promotion decisions. *Journal of Higher Education* 55: 583–589.

Locke, E.A., and Latham, G.P. (1990). *A Theory of Goal Setting and Task Performance*. Englewood Cliffs, NJ: Prentice Hall.

Marsh, H.W. (1976). *The Relationship between Background Variables and Students' Evaluations of Instructional Quality*. OIS 76–9. Los Angeles, CA: Office of Institutional Studies, University of Southern California.

Marsh, H.W. (1982a). Factors affecting students' evaluations of the same course taught by the same instructor on different occasions. *American Educational Research Journal* 19: 485–497.

Marsh, H.W. (1982b). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology* 52: 77–95.

Marsh, H.W. (1982c). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology* 74: 264–279.

Marsh, H.W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology* 75: 150–166.

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

Marsh, H.W. (1985). Students as evaluators of teaching. In T. Husen and T.N. Postlethwaite (eds.), *International Encyclopedia of Education: Research and Studies*. Oxford: Pergamon Press.

Marsh, H.W. (1986). Applicability paradigm: Students' evaluations of teaching effectiveness in different countries. *Journal of Educational Psychology* 78: 465–473.

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research* 11: 253–388. (Whole Issue No. 3)

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

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

Marsh, H.W. (1995). Still weighting for the right criteria to validate student evaluations of teaching in the idea system. *Journal of Educational Psychology* 87: 666–679.

Marsh, H.W. (2001). Distinguishing between good (useful) and bad workload on students' evaluations of teaching. *American Educational Research Journal* 38(1):183–212.

Marsh, H.W., and Bailey, M. (1993). Multidimensionality of students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education* 64: 1–18.

Marsh, H.W., and Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. *Higher Education: Handbook on Theory and Research*(Vol. 8, pp. 143–234). New York: Agathon.

Marsh, H.W., and Dunkin, M.J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R.P. Perry and J.C. Smart (ed.), *Effective Teaching in Higher Education: Research and Practice* (pp. 241–320). New York: Agathon.

Marsh, H.W., Fleiner, H., and Thomas, C.S. (1975). Validity and usefulness of student evaluations of instructional quality. *Journal of Educational Psychology* 67: 833–839.

Marsh, H.W., and Groves, M.A. (1987). Students' evaluations of teaching effectiveness and implicit theories: A critique of Cadwell and Jenkins. *Journal of Educational Psychology* 79: 483–489.

Marsh, H.W., and Hattie, J. (2002). The relationship between research productivity and teaching effectiveness: Complimentary, antagonistic or independent constructs. *Journal of Higher Education* 73: 603–642.

Marsh, H.W., and Hau, K.T. (2003). Big fish little pond effect on academic self-concept: A cross-cultural (26 country) test of the negative effects of academically selective schools. *American Psychologist* 58: 364–376.

Marsh, H.W., and Hocevar, D. (1991a). 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: 9–18.

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

Marsh, H.W., and Overall, J.U. (1979). Long-term stability of students' evaluations. *Research in Higher Education* 10: 139–147.

Marsh, H.W., Overall, J.U., and Kesler, S.P. (1979a). Class size, students' evaluations, and instructional effectiveness. *American Educational Research Journal* 16: 57–70.

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

Marsh, H.W., and Overall, J.U. (1980). Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria. *Journal of Educational Psychology* 72: 468–475.

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

Marsh, H.W., and Roche, L.A. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal* 30: 217–251.

Marsh, H.W., and Roche, L.A. (1994). *The Use of Students' Evaluations of University Teaching to Improve Teaching Effectiveness*. Canberra, ACT: Australian Department of Employment, Education, and Training.

Marsh, H.W., and Roche, L.A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist* 52: 1187–1197.

Marsh, H.W., and Roche, L.A. (2000). Effects of grading leniency and low workloads on students' evaluations of teaching: Popular myth, bias, validity or innocent bystanders? *Journal of Educational Psychology* 92: 202–228.

Marsh, H.W., Rowe, K., and Martin, A. (2002). PhD students' evaluations of research supervision: Issues, complexities and challenges in a nationwide Australian experiment in benchmarking universities. *Journal of Higher Education* 73(3): 313–348.

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

McKeachie, W. (1963). Analysis and investigation of teaching methods. In N.L. Gage (ed.), *Handbook of Research on Teaching* (pp. 448–505). Chicago: Rand McNally.

McKeachie, W.J. (1973). Correlates of students' ratings. In A.L. Sockloff (ed.), *Proceedings: The First Invitational Conference on Faculty Effectiveness Evaluated by Students* (pp. 213–218). Temple University.

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

McKeachie, W.J. (1996). Do we need norms of student ratings to evaluate faculty? *Instructional Evaluation and Faculty Development* 14: 14–17.

McKeachie, W.J. (1997). Student Ratings: The Validity of Use. *American Psychologist* 52: 1218–25.

Murray, H.G. (1980). *Evaluating University Teaching: A Review of Research*. Toronto, Canada, Ontario Confederation of University Faculty Associations.

Murray, H.G. (1983). Low inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology* 71: 856–865.

Murray, H.G. (April, 1987). *Impact of Student Instructions Ratings on Quality of Teaching in Higher Education*. Paper presented at the 1987 Annual Meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 284 495).

Ory, J.C., and Braskamp, L.A. (1981). Faculty perceptions of the quality and usefulness of three types of evaluative information. *Research in Higher Education* 15: 271–282.

Ory, J.C., Braskamp, L.S., and Pieper, D.M. (1980). Congruency of student evaluative information collected by three methods. *Journal of Educational Psychology* 72:321–325.

Overall, J.U., and Marsh, H.W. (1979). Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of Educational Psychology* 71: 856–865.

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

Overall, J.U., and Marsh, H.W. (1982). Students' evaluations of teaching: An update. American Association for Higher Education Bulletin 35(4): 9–13 (ERIC Document Reproduction Services No. ED225473).

Penny, A.R., and Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research* 74(2): 215–253.

Perry, R.P., Abrami, P., Leventhal, L., and Check, J. (1979). Instructor reputation: An expectancy relationship involving student ratings and achievement. *Journal of Educational Psychology* 71: 776–787.

Peterson, C., and Cooper, S. (1980). Teacher evaluation by graded and ungraded students. *Journal of Educational Psychology* 72: 682–685.

Remmers, H.H. (1963). Rating methods in research on teaching. In N.L. Gage (ed.), *Handbook of Research on Teaching* (pp. 329–378). Chicago: Rand McNally.

Renaud, R.D., and Murray, H.G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education* 46: 929–953.

Renaud, R.D., and Murray H.G. (1996). Aging, Personality, and Teaching Effectiveness in Academic Psychologists. *Research in Higher Education* 37: 323–340.

Richardson, J.T.E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment and Evaluation in Higher Education* 30(4): 387–415.

Rindermann, H. (1996). On the quality of students' evaluations of university teaching: An answer to evaluation critique. *Zeitschrift Für Padagogische Psychologie* 10(3–4): 129–145.

Rindermann, H., and Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Research in Higher Education* 42(4): 377–399.

Ryan, J.M., and Harrison, P.D. (1995). The relationship between individual characteristics and overall assessment of teaching effectiveness across different instructional contexts. *Research in Higher Education* 36: 577–594.

Salthouse, T.A., McKeachie, W.J., and Lin, Y.G. (1978). An experimental investigation of factors affecting university promotion decisions. *Journal of Higher Education* 49: 177–183.

Scriven, M. (1981). Summative Teacher Evaluation, in J. Millman (ed.), *Handbook of Teacher Evaluation* (pp. 244–71). Beverley Hills, CA: SAGE.

Sullivan, A.M., and Skanes, G.R. (1974). Validity of student evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology* 66(4): 584–590.

Ting, K. (2000). Cross-level effects of class characteristics on students' perceptions of teaching quality. *Journal of Educational Psychology* 92: 818–825.

Toland, M.D., and De Ayala, R.J. (2005). A Multilevel Factor Analysis of Students' Evaluations of Teaching. *Educational and Psychological Measurement* 65: 272–296.

Watkins, D. (1994). Student evaluations of teaching effectiveness: A Cross-cultural perspective. *Research in Higher Education* 35: 251–266.

Wendorf, C.A., and Alexander, S. (2005). The influence of individual- and class-level fairness-related perceptions on student satisfaction. *Contemporary Educational Psychology* 30: 190–206.

Wilson, R.C. (1986). Improving faculty teaching: Effective use of student evaluations and consultants. *Journal of Higher Education* 57: 196–211.