Making Students' Evaluations of Teaching Effectiveness Effective

The Critical Issues of Validity, Bias, and Utility

Herbert W. Marsh and Lawrence A. Roche
University of Western Sydney, Macarthur

This article reviews research indicating that, under appropriate conditions, students' evaluations of teaching (SETs) are (a) multidimensional; (b) reliable and stable; (c) primarily a function of the instructor who teaches a course rather than the course that is taught; (d) relatively valid against a variety of indicators of effective teaching; (e) relatively unaffected by a variety of variables hypothesized as potential biases (e.g., grading leniency, class size, workload, prior subject interest); and (f) useful in improving teaching effectiveness when SETS are coupled with appropriate consultation. The authors recommend rejecting a narrow criterion-related approach to validity and adopting a broad construct-validation approach, recognizing that effective teaching and SETs that reflect teaching effectiveness are multidimensional; no single criterion of effective teaching is sufficient; and tentative interpretations of relations with validity criteria and potential biases should be evaluated critically in different contexts, in relation to multiple criteria of effective teaching, theory, and existing knowledge.

Heated debate concerning the merits and the shortcomings of students' evaluations of teaching (SETs) continues to flourish, despite intensive ongoing research and international growth in their use as one indicator of teaching quality (Centra, 1993; Feldman, 1997; Marsh & Roche, 1994; Watkins, 1994). In this article, we emphasize the importance of recognizing the multidimensionality of teaching and SETs in understanding research evidence in relation to the validity, perceived bias, and usefulness of SETs. This perspective is important for administrators, program developers, and potential users in making informed decisions regarding the appropriate use of SETs and for future SET research.

Multidimensionality of Teaching

Researchers and practitioners (e.g., Abrami & d'Apollonia, 1991; Cashin & Downey, 1992; Feldman, 1997; Marsh & Roche, 1993) agree that teaching is a complex activity consisting of multiple dimensions (e.g., clarity, teachers' interactions with students, organization, enthusiasm) and that formative–diagnostic evaluations of teachers should reflect this multidimensionality (e.g., a teacher is organized but lacks 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 the usefulness of SET information depend on the content and the coverage of the items. Poorly worded or inappropriate items will not provide useful information, whereas 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, 1987). Valid measurement, however, requires a continual interplay between theory, research, and practice. Several theoretically defensible instruments with a well-defined factor structure have been reviewed (Centra, 1993; Marsh, 1987), but few have been evaluated extensively in terms of potential biases and validity.

The strongest support for the multidimensionality of SETs is based on the nine-factor (Learning/Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Rapport, Breadth of Coverage, Examinations/Grading, Assignments/Readings, and Workload/Difficulty) Students' Evaluation of Educational Quality (SEEQ) instrument (Marsh, 1983, 1984, 1987; Marsh & Dunkin, 1992). These factors are based on various sources (e.g., reviews of current instruments, interviews with students and teachers) and psychometric analyses and were supported by Marsh and Dunkin's evaluation in relation to theories of teaching and learning. Factor analytic support is particularly strong in that more than 30 published factor analyses have consistently identified the nine a priori factors across diverse settings (Marsh, 1987). For example, factor analyses of responses by 50,000 classes (representing responses to nearly one mil-

Herbert W. Marsh and Lawrence A. Roche, Faculty of Education, University of Western Sydney, Macarthur, Campbelltown, New South Wales, Australia.

We thank Alexander Yeung as well as the other authors in this *Current Issues* section for their helpful and provocative interchange.

Correspondence concerning this article should be addressed to Herbert W. Marsh, Faculty of Education, University of Western Sydney, Macarthur, P.O. Box 555, Campbelltown, New South Wales 2560, Australia. Electronic mail may be sent via Internet to h.marsh@uws.edu.au.

 0003-066X/97/$2.00
Vol. 52, No. 11, 1187–1197

lion SEEQ surveys) provided clear support for the SEEQ factor structure on the basis of the total group and on the basis of each of 21 separate subgroups representing different academic disciplines and levels of instruction (Marsh & Hocevar, 1991a). The applicability of the SEEQ to diverse settings in many different countries was investigated in studies reviewed by Watkins (1994), who concluded that "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" (p. 262).

Global or overall ratings cannot adequately represent the multidimensionality of teaching. They also may be more susceptible to context, mood, and other potential biases than are 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 between Abrami and Marsh (and their colleagues), Abrami and d'Apollonia (1991; Abrami, d'Apollonia, & Rosenfield, 1997) seemed to initially prefer the sole use of global ratings for personnel decisions, whereas Marsh (1991, 1994a) preferred a profile of scores, including the different SEEQ factors, global ratings, expected grades, and prior subject interest ratings, but they apparently agreed 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, 1994a). 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 teachers' input into personnel decisions), it might be useful to weight SET factors according to their importance in a specific teaching context. Unresolved issues concerning the validity and the utility of importance-weighted averages (e.g., Marsh, 1994b, 1995), however, dictate caution in pursuing this suggestion. Continuing this debate, Abrami et al.

- raised many concerns about factor analysis that were largely addressed by Marsh (1991, 1994a);
- cited Cashin and Downey (1992) as showing that specific ratings add little to global ratings, but Marsh's (1994b) reanalysis of this study showed that the optimal subset of SETs (in relation to their outcome variable of students' progress ratings) did not even include global items; and
- implied that specific items were less valid than global ratings in multisection validity studies, even though Feldman (1997) reported nine SET dimensions (.57 for organization, .56 for clarity, .46 for impact, .38 for interest stimulation, .36 for discussion, .36 for availability, .35 for elocution, .35 for objectives, and .34 for knowledge) that were more highly correlated with achievement than the .32 correlation between achievement and global ratings reported by Abrami et al.

Abrami et al. reported empirical support for 35 different first-order SET factors (and multiple items representing each factor) that could be represented by 4 higher order factors, prompting Marsh (1994a) to dub the authors as the new champions of the multidimensional perspective.

Many SET instruments fail to provide a comprehensive evaluation of theoretically sound, multiple dimensions of teaching quality, thus undermining their usefulness, particularly for diagnostic feedback. "Homemade" SET surveys constructed by lecturers or committees are rarely evaluated in relation to rigorous psychometric considerations and revised accordingly. This variation in quality also complicates the interpretation and the generalizability of SET research. SET instruments should be designed to measure separate components of teaching effectiveness, and support for the construct validity of the multiple dimensions should be evaluated. The failure to recognize this multidimensionality of SETs undermines the ability to understand their reliability, validity, relation to potential biases, and utility in improving teaching effectiveness.

Reliability, Stability, and Generalizability

The reliability of SETs is most appropriately determined from studies of interrater agreement that assess agreement among different students within the same course (Gillmore, Kane, & Naccarato, 1978; Marsh, 1987). 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 on the number of students rating the class: .95 for 50 students, .90 for 25 students, .74 for 10 students, and .60 for 5 students. Given a sufficient number of students in any one class (or, perhaps, averaged across different classes), the reliability of class-average SETs compares favorably with that of the best objective tests.

Cross-sectional studies typically report that SETs are negatively related to age and years of teaching experience (Feldman, 1983), although SETs may increase slightly during the first few years of teaching. However, in a longitudinal study of changes in ratings of a diverse sample of 195 teachers who had been evaluated continuously over a 13-year period, Marsh and Hocevar (1991b) found no systematic changes in any of the SEEQ factors.

Cross-sectional studies also have shown good agreement between responses by current and former students (see Centra, 1979, 1993; Marsh, 1987). In a 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.

In trying to separate the effects of the teacher and the course on SETs, Marsh (1987) reported that the correlation between overall ratings of different instructors teaching the same course (i.e., a course effect) was −.05, whereas correlations for the same instructor in different courses (r = .61) and in two different offerings of the same course (r = .72) were much larger. These results support the validity of SETs as a measure of teacher effectiveness but not as a measure of the course effective-

ness that is independent of the teacher. Marsh and Bailey (1993) further demonstrated that each teacher had a characteristic profile on the nine SEEQ scores (e.g., high on Organization/Clarity but low on Enthusiasm) that was distinct from the profiles of other instructors and generalized across course offerings over a 13-year period. Such instructor-specific profiles have important implications for the use of SETs as feedback and for their relation to other criteria such as students' learning. The results also provide further support for the multidimensionality of SETs and the generalizability of multidimensional profiles.

Gillmore et al. (1978), applying generalizability theory to SETs, suggested that ratings for a given instructor should be averaged across different courses to enhance generalizability—as many as possible for personnel decisions (they suggested at least five). These recommendations require maintenance of a longitudinal SET archive that would provide the basis for more generalizable summaries, the assessment of changes over time, the determination of which particular courses are best taught by a specific instructor, and more informed personnel decisions.

Validity

SETs are difficult to validate because no single criterion of effective teaching is sufficient (Marsh, 1987, 1994b, 1995). Historically, researchers have emphasized a narrow definition of students' learning—typically operationalized by performance on multiple-choice tests in multisection validity studies (see discussion in the *Students' Learning: The Multisection Validity Study* section)—as the only criterion of effective teaching. We categorically reject the appropriateness of this narrow criterion-related approach to validity that inhibits a better understanding of what SETs measure, of other important criteria of effective teaching, and of the development of a common framework. Marsh (1987) advocated an alternative construct-validation approach in which SETs are posited to be positively related to a wide variety of other indicators of effective teaching and specific SET factors are posited to be most highly correlated with variables to which they are most logically and theoretically related. Within this 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 and former students supports the validity of SETs. The most widely accepted criteria of effective teaching involve students' learning (which can be measured in quite different ways), but many other criteria, such as those considered here, should be studied. A construct-validity approach to the study of SETs now appears to be widely accepted (e.g., Cashin, 1988; Howard, Conway, & Maxwell, 1985) but requires reliable criterion measures that validly reflect effective teaching. Unreliable or invalid criterion measures should not be used as indicators of effective teaching for research, policy formation, feedback to faculty, or personnel decisions.

Students' Learning: The Multisection Validity Study

Students' learning, even when inferred from standardized examinations, typically cannot be compared across different courses. In multisection validity studies (in which multiple sections of the same course are taught by different teachers and evaluated with the same final examination), however, it may be valid to compare teachers in terms of operationally defined learning that can be related to SETs. Despite methodological problems (Marsh, 1987; Marsh & Dunkin, 1992), meta-analyses of multisection validity studies have demonstrated that the sections with the highest SETs are also the sections that perform best on standardized final examinations. Cohen (1987), in his summary of 41 well-designed studies, reported that the mean correlations between students' achievement and different SET components were .55 for structure, .52 for interaction, .50 for skill, .49 for overall course, .45 for overall instructor, .39 for learning, .32 for rapport, .30 for evaluation, .28 for feedback, .15 for interest/motivation, and −.04 for difficulty, in which all but the last two were statistically significant. Validity coefficients tend to be higher for some more specific SET components (Feldman, 1989a, 1997) and for multi-item scales instead of single items (Cohen, 1987). This research demonstrates that SETs reflect students' learning.

Evaluations of Teaching Effectiveness by Different Evaluators

Teaching effectiveness can be evaluated by current students, former students, the teacher himself or herself, colleagues, administrators, or trained observers. Teachers' self-evaluations are useful because they can be collected in all educational settings, are likely to be persuasive for at least the teachers evaluating their own teaching, may be important in interventions designed to improve teaching, and provide insight into how teachers view their own teaching. Feldman's (1989b) meta-analysis of correlations between SETS and self-evaluations reported mean correlations between .15 and .42 for specific SET components and a mean correlation of .29 for overall ratings. In two studies with large numbers of teachers and students evaluating teaching with the SEEQ (Marsh, 1987), (a) separate factor analyses of SETs and self-evaluations identified the same SEEQ factors, (b) student–teacher agreement on every dimension was significant (median rs of .49 and .45 between SETs and teachers' self-evaluations in the two studies) and was typically larger than agreement on overall teaching effectiveness (r = .32), (c) mean differences between students' and faculty members' responses were small, and (d) multitrait–multimethod analyses supported the convergent validity and the discriminant validity of the multidimensional ratings.

Colleagues' and administrators' ratings based on classroom visitations are not very reliable (i.e., ratings by different peers do not even agree with each other) and are not systematically correlated with SETs or other indicators of effective teaching (see Centra, 1979; Koon & Murray, 1996; Marsh, 1987; Murray, 1980).

However, trained external observers may accurately rate some specific classroom teaching behaviors (Marsh, 1987; Murray, 1980). For example, Cranton and Hillgarten (1981) examined relationships between SETs and specific teaching behaviors observed during videotaped lectures in a naturalistic setting: SETs of effectiveness of discussion were higher "when professors praised student behavior, asked questions and clarified or elaborated student responses" (p. 73), and SETs of organization were higher "when instructors spent time structuring classes and explaining relationships" (p. 73). Murray (1983) found that total reports based on 18–24 observations per teacher clearly differentiated between teachers who had previously received high, medium, and low SETs. The average observation reports for each teacher were reliable (even though responses by a single observer were not), and factor analysis of the observations resulted in nine factors like those found in SETs (e.g., Clarity, Enthusiasm, Interaction, Rapport, Organization). These studies show that SETs are logically related to observable teaching behaviors.

Howard et al. (1985; also see Feldman, 1989a, 1989b; Marsh & Dunkin, 1992) compared evaluations by current students, former students, a colleague, and eight trained observers. They concluded that "former-student 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), colleagues' and observers' ratings were not significantly correlated with each other, current SETs, or self-evaluations.

Future Directions

There is no adequate single indicator of effective teaching. 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 ratings by former students, students' achievement in multisection validity studies, teachers' self-evaluations, and extensive observations of trained observers on specific processes such as teachers' clarity. This pattern of results supports their construct validity. Marsh (1987; Marsh & Dunkin, 1992) also discussed the validity of SETs in relation to other important outcomes such as students' motivation, affective criteria, subsequent course-work selection, students' study strategies, and the quality of students' learning. In contrast, research productivity (Hattie & Marsh, 1996) and peers' ratings based on classroom visitations are not systematically related to SETs or other indicators of effective teaching, calling into question their validity as measures of effective teaching. Although most researchers agree that it is necessary to have multiple indicators of effective teaching—particularly for personnel decisions—it is 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.

Within the construct-validity approach, it is important to relate SETs to a wide variety of criteria of effective teaching. For example, there is too little research relating multidimensional SETs to important student outcomes such as motivation, self-concept, affect, metacognition, study strategies, course-work selection, career aspirations, and so forth. Another important element of construct validation for SETs is the need to relate SETs to actual classroom processes. Experimentally manipulated teaching situations provide an underutilized but potentially powerful research tool in this area (e.g., Marsh & Ware, 1982; described below in *The Dr. Fox Effect* section).

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 be fueled or fooled by) SET myths (Feldman, 1997; Marsh, 1987). 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) negligence of the multivariate nature of SETs and potential biases; (d) inappropriate operational definitions of bias and potential biasing variables; and (e) inappropriate experimental manipulations.

Support for a bias hypothesis, as with the study of validity, must be based on a construct-validation approach. If a potential bias (e.g., class size) has a similar influence on multiple indicators of teaching effectiveness (e.g., SETs, self-evaluations, test scores), then the effect may reflect a valid influence on teaching effectiveness. Similarly, if the pattern of relations between a particular background variable and multiple dimensions of SET matches a priori predictions, then the results may support the construct validity of the SETs instead of a bias. For example, the SEEQ factors most logically related to class size are Group Interaction and Individual Rapport. Empirical results indicate that class size is moderately correlated with these two SEEQ factors and nearly uncorrelated with other SEEQ factors (or even positively related to Organization/Clarity) and that a similar pattern is observed in instructors' self-evaluations of their own teaching. These results suggest that class size actually does affect Group Interaction and Individual Rapport in a manner that is accurately reflected in SETs and instructors' self-evaluations, supporting the construct validity of SETs in relation to class size—not a class-size "bias" in SETs.

Marsh (1987; also see Centra, 1979) reviewed several large studies of the multivariate relationship between a comprehensive set of background characteristics and

SETs. In two such studies, 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. However, even these relatively modest relations apparently did not reflect biases. The Workload/Difficulty relation was in the opposite direction than that predicted by a bias (SETs were higher—not lower—in more difficult classes; SETs were lower in "Mickey Mouse" courses). Prior subject interest primarily influenced ratings of Learning/Value and overall course ratings, and a similar pattern of relations was found with teachers' self-evaluations. The most contentious relation, perhaps, was the expected-grade effect, which we consider next.

Expected Grades

What is the size of the relation between class-average expected grades and SETs? Marsh (1987) argued that the class average is the appropriate unit of analysis, reporting correlations between class-average SETs and expected grades varying from −.02 (Breadth of Coverage) to .29 (Learning/Value), .31 (Group Interaction), and .22 and .20 for overall course and teacher ratings. The higher correlation with Learning/Value (also observed with teachers' self-evaluations) is predictable because expected grades reflect, in part, a measure of learning, whereas the higher Group Interaction relation may reflect higher grades in advanced-level seminar courses that facilitate student–teacher interaction. Correlations for global ratings are consistent with the extensive review of this relation reported by Feldman (1976). The single best estimate (based on 9,194 class-average responses from a diversity of different universities, courses, settings, and situations) is probably the .20 value reported by Centra and Creech (1976). More recently, Feldman (1997) concluded that correlations are usually between .10 and .30. Hence, the best estimate of the size of the relation is probably about .20 and certainly no higher than .30.

There are at least three very different interpretations of this relation (Marsh, 1987) and some support for each. First, the grading-leniency hypothesis proposes that instructors who give higher-than-deserved grades will be rewarded with higher-than-deserved SETs, which constitutes a serious bias to SETs. According to this hypothesis, it is not expected grades per se that influence SETs but rather the teacher's leniency in assigning grades. Second, the validity hypothesis proposes that better expected grades reflect better learning by students and that a positive correlation between students' learning and SETs supports the validity of SETs. Third, the students' characteristics hypothesis proposes that preexisting student variables such as prior subject interest may affect students' learning, students' grades, and teaching effectiveness, so that the expected-grade effect is spurious. Although these and related explanations of the expected-grade effect have quite different implications, grades must surely reflect some combination of students' learning, the instructor's grading standards, and students' characteristics.

Multisection validity studies. In these studies (reviewed earlier), sections of a multisection course that performed best on a standardized final examination also gave the most favorable SETs. Because preexisting 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 tend to be as large or larger than the typical expected-grade correlation, it seems that much of this relationship reflects the valid effects of students' learning on SETs. This research provides the strongest basis for the interpretation of the expected-grade effect of any research considered here.

Experimental field studies. Marsh (1984, 1987; Marsh & Dunkin, 1992; also see Abrami, Dickens, Perry, & Leventhal, 1980; Howard & Maxwell, 1982) reviewed experimental field studies purporting to demonstrate a grading-leniency effect on SETs but concluded that the research was weak and flawed. In marked contrast, Haskell (1997) summarized work implying that these studies provide good evidence for a grading-leniency effect, even suggesting an implicit collusion among SET researchers to hide this conclusion. It is important to counter such dubious but popular interpretations, because the use of deception in these studies would presumably fail to meet current ethical standards, making the studies difficult to replicate or refine. Here, we briefly elaborate four crippling weaknesses of these studies by Chacko (1983), Holmes (1972), Powell (1977), Vasta and Sarmiento (1979), and Worthington and Wong (1979), and one subsequent study by Blunt (1991).

The first weakness relates to the ambiguity of deception research. The use of deception as applied in these studies (e.g., reporting false course grades to students, thereby violating students' reasonable grade expectations, sometimes quite seriously) is not only ethically dubious but also methodologically suspect and, as with other deception research (see Lawson, 1997), is unlikely to produce unambiguous, generalizable results.

The second weakness relates to design. In all six of the aforementioned studies, the researchers themselves taught classes in which the students from one large class were randomly assigned to different "grading" groups (Blunt, 1991; Holmes, 1972; Worthington & Wong, 1979) or intact (supposedly equivalent) sections of the same class were graded differently (Chacko, 1983; Powell, 1977; Vasta & Sarmiento, 1979). Hence, there was limited generalizability in all of the studies, potential researcher expectancy effects in three studies, and true random assignment in only three studies.

The third weakness relates to grading-leniency manipulations. These manipulations do not reflect typical,

naturally occurring leniency variation, where students have reasonably accurate grade expectations based on feedback about their actual performance but have not yet received their final grades when completing SETs. In some studies, the experimental manipulation intentionally violated reasonable grade expectations based on actual performance (Blunt, 1991; Holmes, 1972), grades that had no relation to students' actual performance were assigned completely at random (Worthington & Wong, 1979), or students were likely to be aware that they received grades different from those of other students in the same class performing at the same level (Vasta & Sarmiento, 1979; Worthington & Wong, 1979; and maybe Holmes, 1972; Powell, 1977). In one case (Powell, 1977, Study 1), the manipulation was a change in instructional design that was claimed to affect learning, effort, and actual examination performance (and thus was more than a grading-leniency manipulation) or did not involve a manipulation of grading leniency at all (Powell, 1977, Study 2). The size of the manipulation (e.g., ±one letter grade—a two-letter grade difference) sometimes seemed large relative to typical variation (Blunt, 1991; Holmes, 1972; Worthington & Wong, 1979; and maybe Chacko, 1983). In most cases, the final manipulated grade was presented (and emphasized) immediately before SETs were collected (Blunt, 1991; Chacko, 1983; Holmes, 1972; Worthington & Wong, 1979), enhancing the saliency of the grades or violations of reasonable grade expectations. In actual practice, most teachers assign grades that are consistent with expectations they have given students and are monotonic with actual performance according to standards known by students; they do not intentionally mislead students into thinking that they will receive higher or lower grades than they actually receive. Overall, the manipulations seem quite inappropriate.

The fourth weakness relates to results. The published results typically do not provide adequate information to compute effect sizes in any straightforward manner, but the proportions of statistically significant differences between groups (for different SET items) suggest the effects were weak: 2/10 (Blunt, 1991), 5/19 (Holmes, 1972), and 7/50 (Vasta & Sarmiento, 1979; apparently based—inappropriately—on one-tailed tests and on one significant effect in the opposite direction). Worthington and Wong (1979) reported 10/23 significant differences between students who were randomly assigned "satisfactory" and "poor" grades, but they reported only 1/23 significant differences between students assigned "good" and "satisfactory or poor" grades (suggesting a grading-strictness effect but no grading-leniency effect). Chacko (1983) and Powell (1977) reported no significance tests of between-groups differences (although Chacko did report before–after tests for each group separately). In many cases, significant (or apparently largest) differences were for items related to grades and grading fairness (Holmes, 1972; Powell, 1977; Vasta & Sarmiento, 1979; Worthington & Wong, 1979), which might logically be expected to be lower, given the nature of the manipulation.

In summary, this set of experimental field studies is methodologically weak, ethically indefensible, unrepresentative of naturally occurring differences in grading leniency (to the extent that manipulations represent grading leniency at all), and weak in terms of the results. To illustrate the unrepresentative nature of these studies, note that there are likely to be large differences between (manipulated) assigned grades and expected grades in these studies, whereas in practice, expected and actual grades are very similar. In summary, suggestions that this research supports a grading-leniency bias are unwarranted.

Laboratory studies. 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 *The Dr. Fox Effect* section below). Groups of students viewed a videotaped lecture, rated teachers' effectiveness, and completed an examination. When the students returned two weeks later, they were given their examination results and a grade based on their actual performance but scaled according to different standards (i.e., an "average" grade earning a *B*, a *C*+, or a *C*). Students then viewed a similar videotaped lecture, again evaluated teachers' effectiveness, and were tested again. The grading-leniency manipulation had no effect on achievement and weak, inconsistent effects on SETs, failing to support a grading-leniency interpretation.

Other nonexperimental 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. In addressing the issue of how much of the remaining expected-grade effect can be attributed to grading leniency, Howard and Maxwell (1980, 1982) found that most of the covariation between expected grades and SETs was eliminated by controlling for students' prior motivation and students' progress ratings—an indicator of students' learning—suggesting almost no variance due to grading leniency.

In one of the few studies to directly measure grading leniency, Marsh (1987) reported that correlations between teachers' self-perceptions of their own grading leniency (rated on a scale ranging from *easy/lenient grader* to *hard/strict grader*) and both students' and teachers' evaluations of effective teaching were small (*r*s ranged from −.16 to .19), except for ratings of Workload/Difficulty (*r*s of .26 for students' ratings and .28 for teachers' ratings) and teachers' self-evaluations of Examinations/Grading ($r = .32$). On the basis of a separate study, Marsh (1987) reported that self-reported 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.

Summary of expected-grade effects. In summary, evidence from a variety of different studies clearly supports the validity and students' 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 unsubstantial.

In future grading-leniency research, theoretically defensible operational definitions must be developed. Thus, for example, grading leniency seems to be an attribute of the teacher—not individual students within a class—so that relations should be based on class-average results; correlations based on grades and individual SETs within a class seem irrelevant to grading-leniency effects. Furthermore, even class-average expected grades provide only a weak, indirect indicator of grading leniency; more direct measures are required. Studies that assume that high class-average grades reflect grading leniency should be interpreted cautiously. Experimental field studies that have manipulated grading leniency appear to be of limited usefulness because of basic design flaws as well as the methodological and ethical shortcomings of deception-based research, although laboratory experimental studies like those by Abrami et al. (1980) seem more promising. Simple correlational studies seem to be of limited usefulness, but path-analytic approaches are more promising—depending on the variables included. We also find it curious that expected-grade effects are not discussed in relation to multisection validity studies, where students' learning is consistently correlated with SETs in a setting where background characteristics and grading-leniency effects are largely controlled. In this highly regarded design, there is clear evidence for the validity hypothesis that is not contaminated with grading-leniency effects, and the sizes of these effects are typically as large or larger than expected-grade effects reported elsewhere. More generally, this seems like an ideal setting in which to blend qualitative research techniques (on the nature of expected grades and grading leniency) and quantitative techniques that have largely dominated SET research.

The Dr. Fox Effect

The Dr. Fox effect is defined as the overriding influence of instructors' expressiveness on SETs and has been interpreted to mean that enthusiastic lecturers can "seduce" students into giving favorable evaluations, even though the lectures may be devoid of meaningful content. The original Dr. Fox study, as noted by its authors and critics alike, was fraught with methodological weaknesses (see Abrami, Leventhal, & Perry, 1982; Marsh, 1987; Marsh & Ware, 1982), but the original results were seized by some as support for the invalidity of SETs. To overcome some methodological problems, the standard Dr. Fox paradigm was developed, where 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 presented by the same actor (who was called Dr. Fox). Students viewed one of the six lectures, evaluated teaching effectiveness on a multidimensional SET instrument, and completed an achievement test based on all the teaching points in the high-content lecture. Abrami et al.'s (1982) meta-analysis 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 SETs.

In reanalyses 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 affected only Instructor Enthusiasm, the factor most logically related to that manipulation and (b) content coverage significantly affected Instructor Knowledge and Organization/Clarity, the factors most logically related to that manipulation. When students were given no incentives to perform well, instructor expressiveness had more impact on all five SET factors (although the effect on Instructor Enthusiasm was still the largest), but expressiveness also had more impact on 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 in other studies of potential biases, this reanalysis indicates the importance of the multidimensionality of SETs. An effect that 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.

Summary of Potential Bias Interpretations

A summary of typical relationships between background characteristics and students' ratings based on the work of many researchers (see reviews by Marsh, 1987; Marsh & Dunkin, 1992) is presented in Table 1. Whereas a comprehensive review of potential biases is beyond the scope of this article, perhaps the best summary is McKeachie's (1979) conclusion that a wide variety of variables that could potentially influence SETs apparently have little effect.

Utility of Students' Evaluations of Teaching: Improving Teaching Quality

There is substantial testimonial evidence as well as experimental evidence supporting the usefulness of SETs (Centra, 1993; Marsh & Dunkin, 1992; Marsh & Roche, 1994). In most SET feedback studies, teachers are randomly assigned to experimental (feedback) or one or more control groups; SETs are collected during the term; SETs of the teachers in the feedback group are quickly returned to instructors; and the various groups are compared on subsequent SETs and, sometimes, other criterion variables. In Cohen's (1980) meta-analysis, instructors in feedback groups were subsequently rated 0.30 standard deviations higher than controls on a total rating, and even larger differences were observed for ratings of Instructor Skill, Attitude Toward Subject, and Feedback to Students. Studies that augmented feedback with consultation pro-

Table 1
Overview of Relationships Found Between Students' Ratings and Background Characteristics

Background characteristic	Summary of findings
Prior subject interest	Classes with higher interest rate classes more favorably, although it is not always clear if interest existed before the start of the course or was generated by the course or the instructor.
Expected grade–actual grade	Class-average grades are correlated with class-average students' evaluations of teaching, but the interpretation depends on whether higher grades represent grading leniency, superior learning, or preexisting differences.
Reason for taking a course	Elective courses and those with a higher percentage of students taking the 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 are rated somewhat more favorably, although some find curvilinear relationships where large classes also are rated favorably.
Level of course or year in school	Graduate-level courses are rated somewhat more favorably; weak, inconsistent findings suggest upper division courses are rated higher than lower division courses.
Instructor's rank	Mixed findings but little or no effect.
Sex of instructor 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 ratings are known to be used for tenure–promotion decisions.
Administrative conditions	Somewhat higher if ratings are not anonymous and the instructor is present when ratings are being completed.
Students' personality	Mixed findings but apparently little effect, particularly because different "personality types" may appear in somewhat similar numbers in different classes.

Note. Particularly for the more widely studied characteristics, some studies have found little or no relation or even results opposite to those reported here. The size, or even the direction, of relations may vary considerably, depending on the particular component of students' ratings that is being considered. Few studies have found any of these characteristics to be correlated more than .30 with class-average students' ratings, and most relations are much smaller.

duced substantially larger differences, but other methodological variations had little effect. Overall and Marsh (1979) also showed that feedback with consultation led to improved examination performance and affective outcomes as well as higher SETs. The most robust finding from this research is that consultation augments the SET effects, but there is insufficient information about the most effective type of consultative feedback.

The use of norms helps teachers to determine their relative strengths and weaknesses, because raw scores on different factors are not directly comparable (e.g., 3.5 on a 5-point scale for Organization/Clarity does not indicate whether Organization/Clarity is good, and 3.75 on Enthusiasm may not be better than 3.5 on Organization/Clarity). If, however, the mean Organization/Clarity rating in similar courses is 4.0, then a 3.5 suggests that a particular teacher may need to improve his or her organization. Also, the presentation of graphic profiles emphasizes a teacher's relative strengths and weaknesses across different dimensions rather than comparisons with other teachers, a feature that may offset any demotivation produced by receiving below-average SETs.

Marsh and Roche (1993) reviewed different forms of feedback, evaluated a feedback–consultation intervention adapted from Wilson (1986), and compared the effects of midterm and end-of-term feedback. Teachers completed self-evaluations and were evaluated by students at the middle of Semester 1 and at the end of Semesters 1 and 2. Three randomly assigned groups received the intervention at the midterm of Semester 1, at the end of Semester 1, or not at all (control). A key component 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 strategies for that factor. Ratings for all groups improved, but improvement was significantly greater in the intervention groups than in 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 (as compared with the control group), 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. It is important to note that this intervention can be conducted only with a well-designed, multidimensional instrument like the SEEQ and that the specificity of the intervention effects to the targeted dimensions further supports the construct validity of multidimensional SETs.

Conclusions and Implications

Confusion about the validity and the effectiveness of SETs will continue as long as the various distinct components of students' ratings are treated as a single "puree" rather than as the "apples and oranges" that make up effective teaching. In evaluating the validity and the usefulness of SETs, practitioners and researchers are encouraged to consider whether proper account has been taken of the distinct components of students' ratings that reflect the multidimensionality of effective teaching. SET instruments differ markedly in their ability to measure these distinct components, and as Abrami et al. (1997) showed, no amount of factor analytic gymnastics is able to salvage apples and oranges from pureed questions. Multidimensionality is important not only because of its obvious diagnostic utility as instructor feedback but also because it provides a more sophisticated and realistic assessment of the various aspects of teaching. Thus, various contextual variables, possible biasing influences, and validity all can be investigated more systematically and productively, rather than lumping all the different dimensions into a puree and then trying to separate out the causal ingredients!

REFERENCES

Abrami, P. C., & 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., & Rosenfield, S. (1997). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 321–367). New York: Agathon Press.

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

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

Blunt, A. (1991). The effects of anonymity and manipulated grades on student ratings of instructors. *Community College Review, 18,* 48–54.

Cashin, W. E. (1988). *Student ratings of teaching: A summary of research* (IDEA Paper No. 20). Manhattan: Kansas State University, Division of Continuing Education.

Cashin, W. E., & 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: Jossey-Bass.

Centra, J. A. (1993). *Reflective faculty evaluation.* San Francisco: Jossey-Bass.

Centra, J. A., & Creech, F. R. (1976). *The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness* (Project Rep. No. 76-1). Princeton, NJ: Educational Testing Service.

Chacko, T. I. (1983). Student ratings of instruction: A function of grading standards. *Educational Research Quarterly, 8,* 19–25.

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. (1987, April). *A critical analysis and reanalysis of the multisection validity meta-analysis.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

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

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52,* 1198–1208.

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

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. (1989a). 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. (1989b). 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. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368–395). New York: Agathon Press.

Frey, P. W. (1978). A two-dimensional analysis of student ratings of instruction. *Research in Higher Education, 9,* 69–91.

Gillmore, G. M., Kane, M. T., & 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. (1997a). *Do manipulated course grades influence student ratings of instructors? A small meta-analysis.* Manuscript in preparation.

Greenwald, A. G. (1997b). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52,* 1182–1186.

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

Haskell, R. (1997). Academic freedom, tenure and student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis Archives, 5.* Retrieved from World Wide Web: http://olam.ed.asu.edu/epaa/

Hattie, J., & Marsh, H. W. (1996). The relationship between research and teaching in universities. *Review of Educational Research, 66,* 507–542.

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology, 63,* 130–133.

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

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

Koon, J., & Murray, H. G. (1996). Using multiple outcomes to validate student ratings of overall teacher effectiveness. *Journal of Higher Education, 66,* 61–81.

Lawson, E. (1997). Deception research: After 30 years of controversy. In M. Bibby (Ed.), *Ethics and educational research* (pp. 15–48). Coldwater, Victoria, Australia: Australian Association of Research in Education.

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. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future re-

search. *International Journal of Educational Research, 11*(Whole Issue No. 3).
Marsh, H. W. (1991). 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. (1994a). Comments to: "Review of the Dimensionality of Student Ratings of Instruction: I. Introductory Remarks. II. Aggregation of Factor Studies. III. A Meta-Analysis of the Factor Studies." *Instructional Evaluation and Faculty Development, 14,* 13–19.
Marsh, H. W. (1994b). Weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology, 86,* 631–648.
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., & Bailey, M. (1993). Multidimensionality of students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education, 64,* 1–18.
Marsh, H. W., & Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook on theory and research* (Vol. 8, pp. 143–234). New York: Agathon Press.
Marsh, H. W., & 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., & 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., & 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., & Roche, L. A. (1994). *The use of students' evaluations of university teaching to improve teaching effectiveness.* Canberra, Australian Capital Territory, Australia: Department of Employment, Education and Training.
Marsh, H. W., & 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. J. (1979). Student ratings of faculty: A reprise. *Academe, 65,* 384–397.
McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52,* 1218–1225.
Murray, H. G. (1980). *Evaluating university teaching: A review of research.* Toronto, Ontario, 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, 75,* 138–149.
Overall, J. U., & 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., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology, 72,* 321–325.
Powell, R. W. (1977). Grades, learning, and student evaluation of instruction. *Research in Higher Education, 7,* 193–205.
Vasta, R., & Sarmiento, R. F. (1979). Liberal grading improves evaluations but not performance. *Journal of Educational Psychology, 71,* 207–211.
Watkins, D. (1994). Student evaluations of teaching effectiveness: A cross-cultural perspective. *Research in Higher Education, 35,* 251–266.
Wilson, R. C. (1986). Improving faculty teaching: Effective use of student evaluations and consultants. *Journal of Higher Education, 57,* 196–211.
Worthington, A. G., & Wong, P. T. P. (1979). Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology, 71,* 764–775.

Postscript

Greenwald's (1997a) meta-analysis (based on five grade-manipulation studies that we critiqued) raises unresolved concerns for the authors of the other articles in this *Current Issues* section about technical details, studies included and not included, quality and number of studies, and effect-size calculations. On the basis of these concerns, our critical evaluation of these studies described earlier, and our reanalysis of Greenwald's data (graciously provided to us by Greenwald), we disagree with his conclusion that effect sizes are "moderate to large."

D'Apollonia and Abrami's (1997, this issue) ongoing concerns about factor analysis, multitrait–multimethod analysis, and our multidimensional perspective are addressed by Marsh (1994a). Their implication that this perspective is based only on factor analysis is countered by the diversity of multidimensional construct-validation research emphasized in our article.

Greenwald and Gillmore (1997, this issue) pursued the worthy goal of enhancing SET validity by attempting to statistically control the undesirable influence of grading leniency. We concur with other authors in this *Current Issues* section that this ambitious goal was not met, and we outline our most serious concerns.

First, we contend that many of Greenwald and Gillmore's (1997) arguments against what they refer to as the teaching-effectiveness theory (the validity hypothesis) rely on their stated assumption that teaching effectiveness is constant for all students within a class. However, as McKeachie (1997, this issue) explains, what works for one student may not work for others—an adage well-known to teachers and SET researchers.

Second, we contend that each of Greenwald and Gillmore's (1997) data patterns is open to alternative explanations consistent with teaching-effectiveness interpretations. For example, they conclude that higher relative expected grades (getting better-than-usual grades) reflects grading leniency, but better grades also reflect better-than-usual mastery, thus supporting the validity (teaching-effectiveness) hypothesis. McKeachie (1997) offers other examples, and we join d'Apollonia and Abrami (1997) in encouraging readers to generate their own plausible reinterpretations of the results in relation to teaching-effectiveness theory and additional influences not involving grading leniency and to critically evaluate the grading-leniency explanations provided by Greenwald and Gillmore. Therefore, we disagree with Greenwald and Gillmore's conclusion that the grades–ratings correlation is an effect of grading leniency on the basis of such correlational data.

Third, we contend that within a given class, students get different grades mostly because they differ in background, effort, amount learned, and so forth, not because the teacher uses different grading standards for different students within the same class. Furthermore, because grading leniency is a teacher attribute, we contend that analyses should be based on class-average responses, as is widely acknowledged in the SET literature. Thus, within-class correlations involv-

ing expected grades of individual students seem not to reflect grading leniency so that deductions based on Greenwald and Gillmore's (1997) within-class correlations seem largely irrelevant in relation to teacher grading leniency.

Fourth, Greenwald and Gillmore's (1997) critical variable should be grading leniency (not expected grades), and they should ask the following question: If one gives higher than deserved grades, will one get higher than deserved ratings? Whereas it would be desirable to measure grading leniency separately from expected grades, we accept the difficulty of this task (one that should be pursued in further research like our exploratory attempts described earlier). When expected grades are used to infer grading leniency, however, it is important that all reasonable attempts are made to control for effects other than grading leniency (e.g., students' learning and preexisting background differences) and to fully acknowledge the inherent difficulty in this task in interpreting and applying the results.

Greenwald and Gillmore's (1997) research does reveal how much class-average-expected-grade variance can be explained by grading leniency, students' learning, presage variables, and so forth. They conclude, however, that increasing grades by two standard deviations "should produce" a change of one standard deviation in SETs and that "Yes, I can get higher ratings by giving higher grades" (p. 1214). We assert that these conclusions (and their statistical adjustment) inappropriately imply causation from correlation. Their path models should include measures of grading leniency "uncontaminated" by students' learning (or separate measures of the two constructs) and preexisting background variables (e.g., prior subject interest, course level) that may affect expected grades and SETs. Even then, causal inferences would be highly speculative. Moreover, multisection validity studies show that SETs are related to students' learning when grading leniency is controlled. These correlations of .30 to .40 (up to .57 for SET Organization ratings) are larger than typical correlations between SETS and expected grades (.20 to .30), suggesting that any unique variance attributed to grading leniency would be negligible. For these reasons, we conclude that correcting for expected grades (instead of the intended target—grading leniency) actually eliminates valid effects of good teaching reflected in superior learning and higher grades, throwing the validity baby out with the bias bathwater.

We agree with McKeachie (1997) that SETs are multidimensional; broader construct-validity perspectives are needed; SET application for improving teaching effectiveness and personnel decisions needs improvement; and class-size effects are valid, not biased. However, we caution that the Instructional Development and Effectiveness Assessment students' progress ratings lack discriminant validity (Marsh, 1994b, 1995) and have not been validated in relation to traditional SET criteria (e.g., students' learning). In relation to global ratings and norms, we encourage creative strategies for how teachers and administrators can best use all available SET information—perhaps with consultation—rather than denying them access to information, fearing they will misuse it.

Greenwald (1997b, this issue), in his postscript, now acknowledges that his aim was to "upset the dominant view" of recognized SET scholars. His expressed puzzlement with acceptance of the dominant view, we suggest, stems from overreliance on a global view of SETS and criterion-related-validity perspectives. Effective teaching reflected by SETs is only one of many potential influences on students' achievement, and students' achievement is only one of many SET criteria. Hence, the correlation of .32 between global ratings and achievement and correlations of up to .57 for Organization ratings are reasonable. Also, demonstrating the importance of our multidimensional perspective, we showed that enthusiasm and class size do have valid effects. Furthermore, it is important to distinguish unfairness from bias (class size is a valid influence, not a bias, but may or may not be fair). However, Greenwald and Gillmore (1997) have provoked new interest in a field where leading authorities may have become complacent, and for this we thank them.