Journal of Educational Psychology
March 2000 Vol. 92, No. 1, 202-228

Effects of Grading Leniency and Low Workload on Students' Evaluations of Teaching
Popular Myth, Bias, Validity, or Innocent Bystanders?

Herbert W. Marsh

Lawrence A. Roche
Faculty of Education and Languages University of Western Sydney Macarthur

ABSTRACT

Two studies debunk popular myths that student evaluations of teaching (SETs) are substantially biased by low workload and grading leniency. A workload bias is untenable because the workload-SET relation is positive. The small grade-SET relation (.20 for overall ratings) has many well-supported explanations that do not involve bias. Some SET factors (e.g.,Organization, Enthusiasm) are unrelated to grades, and the highest relation is with Learning (.30), implying valid teaching effects rather than bias. Structural equation models confirmed that perceived learning and prior characteristics (course level, prior subject interest) account for much of the grade-SET relation. The relation is also nonlinear, so that high grades (sometimes misused as a leniency measure) are unrelated to SETs. Contrary to dire predictions based on bias claims, Workload, expected grades,and their relations with SETs were stable over 12 years.

Major reviews of the voluminous student evaluation of teaching (SET) literature (Abrami, d'Apollonia, & Cohen, 1990 ; Cashin, 1988 ; Cohen, 1987 ; Feldman, 1989a , 1989b , 1997 , 1998 ; Marsh, 1984 , 1987 ; Marsh & Dunkin, 1992 ; Marsh & Roche, 1994 , 1997 ; McKeachie, 1979) have consistently shown that,with careful attention to measurement and theoretical issues, SETs are multidimensional, reliable, relatively valid in relation to various indicators of teaching effectiveness, useful for teaching improvement, and relatively unaffected by suspected biasing factors such as class size, grading leniency, and workload. Marsh (1987) concluded that SETs are probably "the most thoroughly studied of all forms of personnel evaluation, and one of the best in terms of being supported by empirical research" (p. 369). Despite such impressive support and intensive ongoing research and international growth in the successful use of SETs as one indicator of teaching quality (Marsh, 1986a ; Watkins, 1994),unsubstantiated claims of potential biases in SETs continue to flourish. One particularly pervasive allegation is that to obtain good SETs, teachers need only reduce the workload for students and give undeserved high grades (Greenwald & Gillmore, 1997a , 1997b). An implicit or explicit hypothesis in this assertion is that low workloads and easy grading standards positively bias SETs. The present investigation used a construct validity approach to examine support for these bias hypotheses. To address these issues, we critically review previous research, reanalyze recently published data, and present new analyses based on multidimensional SETs.

Popular Myths: An Anecdotal Approach to Bias

Why do what Aleamoni (1987) and Feldman (1997) refer to as academic myths about biases endure despite clear refutation? Perhaps for each large, representative, well-designed study, there is another study, comment, or electronic bulletin-board message that relies on an atypical anecdote or an appeal to popular myth for its impact. The following less sensational anecdote effectively illustrates some deeper issues at work in the observed relationships between workload, expected grades, and SETs:

> A science lecturer used a well-validated multidimensional SET rating form as diagnostic feedback to improve her teaching. Her enthusiasm and breadth of coverage, for example, were rated highly,but students rated workload (pace and difficulty) as too heavy and learning as rather low. She accepted that these were valid student concerns, but faced considerable departmental pressure to simply let the students "sink or swim." Charges of "dumbing down" emerged when she made significant changes to her course delivery: She set more realistic goals, pitched material at an appropriate level where students could "reach" it, and emphasized a solid understanding of the course basics. Students' motivation, learning, and positive attitude toward the subject matter soon improved, and they performed better than had previous classes on an equivalent final exam. Students recognized her outstanding teaching with higher SETs, and her colleagues, impressed with the quality and enthusiasm of students emerging from her classes, nominated her for a university-wide Outstanding Teaching Award (which was based on a range of criteria that extended beyond student ratings), which she subsequently received.

This case study highlights several themes that underlie our reexamination of bias claims (e.g., Greenwald & Gillmore 1997a , 1997b). First, the multidimensionality of SETs facilitates a constructive diagnosis of relevant issues, whether for teaching improvement or research (Marsh & Roche, 1993 ; 1997). Second, despite consistent findings of an overall positive correlation between workload and SETs (contrary to popular academic mythology, more difficult courses are rated more highly), a curvilinear relation is plausible such that beyond a certain point,excessive workload may reduce teaching effectiveness and SETs. In such cases, reducing course difficulty is likely to enhance student learning and engagement rather than diminish it. Third, good teaching produces many desirable outcomes, including motivation,better learning, and higher expected grades. Expected grades cannot be used as a measure of grading leniency that is unconfounded with actual learning and motivational outcomes.

A Construct Validity Approach to Bias

A growing body of flawed, misleading research appears to fuel and to be fueled by SET myths (see Feldman, 1997 ; Marsh & Roche, 1997). Typical methodological problems include (a) inappropriate operational definitions of bias and potential biasing factors, (b) neglect of the multidimensionality of SETs and other constructs, (c) inappropriate use of the student as the unit of analysis instead of the class average, (d) reliance on small or idiosyncratic samples (particularly when the appropriate sample size is the number of classes), (e) causal interpretations of correlations, and (f) 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, we use a broad construct validity approach, 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.

Here, we demonstrate this construct validity approach to evaluating competing interpretations of the small but contentious relations between SETs, workload, and class-average grade expectations (grades). We begin with a brief overview of SET bias research, focusing particularly on workload-SET and grade-SET relations. We then evaluate Greenwald and Gillmore's (1997a , 1997b ; Greenwald, 1996 , 1997) interpretation of the grade-SET relation as a grading leniency bias with a causal effect on SETs that is best understood in the context of workload ratings. In Study 1, we reanalyze their published data, demonstrating that their data are more consistent with a validity interpretation of SETs than with their bias interpretation. In Study 2, we present new analyses of multiwave,longitudinal data that are based on the multidimensional Students' Evaluations of Educational Quality (SEEQ) instrument to address issues of the valid interpretations of relations between SETs,expected grades, and workload and to test "doom and gloom" implications of bias interpretations.

The multidimensionality of SETs is widely acknowledged in relation to validity and, in particular, to the utility of diagnostic feedback (Feldman, 1997 ; Marsh & Roche,1997), but it is also very important for evaluating bias interpretations. If, for example, a particular variable exhibits reasonably uniform relations with different SET factors when theory or logic dictates that they should differ, then there may be evidence of a bias. However, if the sizes of relations differ systematically for different SET factors and these differences match theoretical or logical a priori predictions, then at least a simple bias interpretation seems untenable. Similarly, a construct validity approach suggests that if a background variable has a similar influence on multiple measures of teaching effectiveness (e.g.,SETs, teacher self-evaluations, student motivation, subsequent course choice, test scores), then the effect may reflect a valid influence on teaching effectiveness rather than a bias. To illustrate this approach, consider the relation between enrollment (class size) and SETs. Two SEEQ factors, Group Interaction and Individual Rapport, logically relate negatively to enrollment. Empirical results confirm that enrollment is moderately negatively correlated with these two SEEQ factors and nearly uncorrelated, or even slightly positively related, with other SEEQ factors, and that a similar pattern is observed in teacher self-evaluations of their own teaching. These results support a priori predictions that enrollment actually does affect Group Interaction and Individual Rapport, as accurately reflected in SETs and instructor self-evaluations, thus supporting the construct validity of SETs in relation to enrollment, not an enrollment bias in SETs. Also, Marsh (1987) suggested that the class-size-SET relation is nonlinear, such that beyond some inflection point, SETs increase with increasing enrollment, a finding that is inconsistent with a simple bias hypothesis. Clearly, the nature of observed relations must be carefully scrutinized before bias interpretations are offered on the basis of correlational results.

In another classic illustration of this approach, Marsh and Ware (1982) demonstrated that experimentally manipulated teacher enthusiasm effects in the "Dr. Fox" studies, which had previously been interpreted as a bias to SETs, were limited primarily to student ratings of teacher Enthusiasm and had much less effect on other SET factors. Because Enthusiasm ratings should be substantially related to the teacher expressiveness manipulation,the results support the construct validity of multidimensional SETs. The importance of Enthusiasm was also supported in that, under certain conditions, enthusiastic teaching led to better subsequent performance on standardized examinations (Marsh, 1987). We now pursue this construct validity approach, emphasizing the multidimensionality of SETs, to evaluate interpretations of SET correlations with workload and expected grades.

Workload: A Bias or an Important Part of Teaching Effectiveness?

Reflecting an apparent ambiguity in research and practice, course workload is viewed as both a legitimate component of teaching effectiveness and a background variable that may be a potential bias to SETs. In factor analyses of SEEQ responses, the four Workload items (difficulty, workload, pace, hours per week outside of class) consistently form a well-defined, distinct factor (e.g., Marsh & Hocevar, 1991). Applicability studies also confirm students' perceptions of workload items as relevant to teaching quality (Marsh, 1986a ; Watkins, 1994). Marsh and Dunkin (1992) provided a theoretical rationale for why Workload is an important aspect of effective teaching. Workload that is seen by students to be far too much or far too difficult is-almost by definition-imposed without due consideration of learners' capacities and prior learning. Similarly, if the pace is too fast,the material is unlikely to be absorbed, and learning suffers. Overloaded students find it difficult to experience subjective feelings of success, receive little or no reinforcement, and may be forced to adopt learning strategies that reduce their ability to understand and generalize from the specific learning context. Conversely, if success is too easily won as a result of an overly light workload, students may lose interest and devalue such learning. Students tend to value learning and achievement more highly when it involves a substantial degree of challenge and commitment (McKeachie, 1997a). This theoretical account predicts a small positive overall relation between Workload and other SET factors and a nonlinear component whereby SETs increase as workload increases to an optimal level, then flatten out or even decline for an excessive workload.

Workload is frequently raised as a potential bias to SETs in the belief that offering easy courses leads to better SETs. Whereas the workload effect was one of the largest in SEEQ research (Marsh, 1987 ; Marsh & Dunkin, 1992), the direction of the effect was opposite to that expected if it was a bias; workload was positively correlated with SETs. Other research reviewed by Marsh (1987) was generally consistent with SEEQ results. Marsh and Overall (1979) also reported that instructor self-evaluations of their teaching effectiveness tended to be positively related to workload. In the applicability paradigm, which has been conducted in universities all over the world, students selected a representative *good* and *poor* teacher and rated each using SEEQ. These results (Marsh, 1986a ; Watkins,1994) consistently demonstrate that good teachers require higher levels of workload than do poor teachers. Overall SETs are also positively related to workload items in the large, multi-institution Instructional Development and Effectiveness Assessment (IDEA) instrument database (Cashin, 1988). Other rating instruments that include a Workload factor typically report that more difficult courses receive somewhat more favorable ratings (e.g., Centra, 1993 ; Centra & Creech, 1976 ; Freedman & Stumpf, 1978 ; Frey, 1978 ; Linn, Centra, & Tucker,1974 ; Michigan State University, Office of Evaluation Services, 1972). Pohlman (1975) also reported significant relations between hours outside of class and student ratings. Schwab (1976) reported a positive effect of perceived difficulty on SETs, even after controlling other background variables, such as grades.

Gillmore and Greenwald (1994) reported significant positive correlations between global SET ratings and three of the four items that they used to infer workload (challenge, .35; effort,.14; involvement, .24; total hours per credit, .03, *ns*). Also, the authors asked students to estimate the total number of hours spent on the class that were useful (as well as the total number of hours). Students perceived most hours to be valuable, and valuable hours were even more highly correlated with global SETs ($r = .62$) than were other workload items. Franklin and Theall (1996) obtained similar results across different disciplines from two different universities. Gillmore and Greenwald (1994) ,acknowledging empirical support for a positive workload-SET relation, pondered the conundrum of why faculty nevertheless feel that courses that are more difficult are rated lower.

In summary, because the direction of the workload effect is opposite to that predicted as a potential bias, and because this finding is consistent for both SETs and instructor self-evaluations,workload does not appear to constitute a bias to SETs. More research is needed, however, to test the suggestion that there might be a nonlinear component in this positive relation. This hypothesis is explored in Study 2 of the present investigation.

Class-Average Grade Expectations: Can Inflated Ratings Be Bought With Inflated Grades?

The small, positive grade-SET relations are probably the most hotly debated topic in the literature on potential SET biases. Briefly, there are at least three competing interpretations of grade relations (Marsh, 1987 ; Marsh & Roche, 1997). First,the *validity hypothesis* proposes that higher grades reflect better student learning and that a positive correlation between student learning and SETs supports the validity of SETs. The strongest support for this interpretation comes from multisection validity studies (discussed later). Second, the *prior characteristics hypothesis* proposes that preexisting student or course variables such as prior subject interest, prior motivation, class size, or course level affect student learning, grades, and actual teaching effectiveness such that the grade relation may be spurious. Third,the *grading leniency hypothesis* proposes that instructors who give higher than deserved grades are rewarded with higher than deserved SETs, constituting 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. Hence, a critical concern is how to either measure grading leniency directly or isolate the component of grades attributable to grading leniency. These and other explanations of the grade-SET relation have quite distinct implications, but actual or expected grades must surely reflect some combination of student learning, the instructor's grading standards, and prior characteristics.

Greenwald and Gillmore (1997a) also proposed an *attribution hypothesis* (see also Perry, 1997 ; Snyder & Clair, 1976 ; Theall, Franklin, & Ludlow, 1990). Within attribution theory, a well-established phenomenon called the *self-serving effect* (e.g., Marsh, 1986b) leads students to internalize responsibility for their successes but externalize responsibility for failure. In the present context, this implies that students tend to attribute high grades to internal characteristics, such as their own ability, effort, or study skills,and to attribute disappointing grades to external characteristics,such as poor teaching, poor quality examinations, course difficulty,or, perhaps, unusually stringent grading standards. Snyder and Clair (1976) offered some support for these proposals in that for experimentally manipulated expected and obtained grades,students obtaining higher than expected grades attributed them more internally (to themselves) but those obtaining lower than expected grades attributed the cause more externally (to the teacher) compared with students who got their expected grade. Attribution theory implies an asymmetry or nonlinearity in predicted grade relations in that teachers might not be rewarded for giving high grades (whether due to good teaching or lenient grading standards) because students may attribute these to internal characteristics,but teachers may be punished for students' low grades (perhaps deservedly, if the grades were due to poor teaching). Also, it is important to emphasize that the self-serving effect might or might not actually reflect a bias, in that if a particularly bright class of students works hard and still receives poor grades, then poor teaching could be a plausible explanation for the students' lack of success. Similar attribution effects in teacher perceptions may help to explain Gillmore and Greenwald's conundrum of why many faculty members feel that harder courses are rated lower despite considerable contrary evidence. Teachers may accept positive SETs but externalize poor SETs by interpreting them as biased or inaccurate or by giving undue weight to dubious studies that support popular myths despite the preponderance of contrary evidence. Although not a major focus of this investigation, we agree with Greenwald and Gillmore's contention that attribution theory may be useful in exploring potential biases to SETs as well as teachers' potentially biased reactions to SETs.

Size of the Grade-SET Relation

There is good agreement, at least, that the grade-SET relation is positive. It is, however, important to establish the size of this relation. Estimates of .20 (Centra & Creech,1976) and .23 (Howard & Maxwell, 1980) for the relation between expected grades and global SETs were reported for two commercially available instruments and were based on two large population-like databases, each of which represented a cross-section of disciplines and U.S. universities. Marsh (1980 , 1983), on the basis of a large sample representing different academic disciplines, also reported correlations between overall teacher ratings and expected grades to be about .20. Feldman (1976 , 1997) reviewed grade-SET relations reported in the SET literature and found that the relation was typically between .10 and .30. Feldman (1997) argued that the popular faculty perception of a high grade-SET correlation is a myth, because the relation is consistently small. He also noted that at least some of this small relation could not be interpreted as bias, because it reflects valid covariation. Thus, the worst-case scenario is that a fraction of the small relation might reflect a bias. In summary, various sources arrive at a remarkably similar conclusion: The grade-SET relation is small, about .20 for global teacher ratings.

Within a construct-validity approach that emphasizes the multidimensionality of SETs, it is important to evaluate how the sizes of grade-SET relations vary for different SET factors. A simple bias hypothesis, for example, might posit that grade-SET relations are similar in size across different SET factors. Marsh (1984 , 1987), however, reported that grade relations with different SEEQ factors varied from close to zero (for Organization and Breadth of Coverage) to about .30 for Learning/Value (consistent with the validity hypothesis that higher grades reflect better learning) and for Group Interaction (consistent with a prior characteristics hypothesis that both grades and Group Interaction tend to be higher in small, specialized seminars and lower in large, introductory courses; controlling the group-interaction-grade relation for enrollment and percentage of students majoring in the department reduced the relation by about one third). In contrast, the grade-workload correlation was negative (-.34). Marsh (1980 , 1984 , 1987) argued that this relation was logically consistent in that classes of students expecting to receive lower grades naturally tend to consider the class to be more difficult, faster paced, having a heavier workload, and requiring more hours outside of class. In this sense, grades are part of the definition of what constitutes a more difficult class (i.e., classes in which everyone gets lower grades are more difficult and, therefore, require more work for a student to do well). Also consistent with this relation is an explanation based on attribution theory, which suggests that students tend to attribute poor grades to external causes such as workload and course difficulty.

Marsh (1984 , 1987) extended the examination of construct validity to relations between expected grades and teacher self-evaluations of their own teaching effectiveness. These correlations tended to be smaller than those based on student ratings, but the pattern of relations with corresponding SET factors was similar: Grades were most positively correlated with teacher self-evaluations of Group Interaction (.17) and Learning/Value (.11) and were most negatively correlated with teacher self-evaluations of their students' Workload (-.19). Patterns of relations between grades, different SET factors, and different teacher self-evaluation factors support the validity and prior characteristics hypotheses,but not simple bias hypotheses.

Multisection Validity Studies

One of the most established findings in the SET literature is that SETs are positively related to objective measures of student achievement, a criterion typically endorsed as the most important criterion of effective teaching (although many other indicators should also be considered). Such findings are based

on the multisection validity paradigm in which multiple sections of the same course are taught by different teachers and evaluated with the same final examination. This design facilitates comparison of teachers' effectiveness in terms of operationally defined learning that can be related to SETs. Despite methodological complications (Abrami et al., 1990 ; Marsh, 1987 ; Marsh & Dunkin, 1992), meta-analyses of multisection validity studies demonstrate that the sections with the highest SETs are also the sections that perform best on standardized final examinations. P. A. Cohen (1987) , in his summary of 41 well-designed studies, reported that the mean correlations between achievement and different SET components were .50 or above for structure,interaction, and skill and .40 or above for overall teacher and course ratings. Validity coefficients were even higher for some more specific SET components (Feldman, 1989a) and for multi-item scales instead of single-item ratings (P. A. Cohen, 1987). This research demonstrates that SETs reflect student learning and is broadly accepted as providing support for the validity of SETs.

Multisection validity studies also provide particularly strong support for the validity interpretation of grade-SET relations. The validity coefficients in these studies are relations between class-average grades and SETs. Because preexisting differences and grading leniency are largely controlled in these studies, the results provide a reasonably pure test of the validity hypothesis. Furthermore, the size of the grade-SET relation (about .45 for overall teacher ratings) in multisection validity studies is larger than the size of the typical grade-SET correlation (about .20). Hence, the valid effect of student achievement is able to explain most of the observed grade-SET relation such that there is little or no variance left to be explained by grading leniency. However, caution must be used in extrapolating these results to a more general setting where the strong grade-SET relation in multisection validity studies may be attenuated when other features are not controlled. Because the multisection validity studies provide irrefutable support for the validity explanation, they also imply that grades cannot be interpreted as the effects of grading leniency unless the effects of true achievement (as well as preexisting student and course characteristics) are controlled.

Grade Manipulation Studies

Marsh (1987 ; Marsh & Dunkin, 1992 ; see also Abrami, Dickens, Perry, & Leventhal, 1980 ; Howard & Maxwell, 1982) reviewed experimental field studies that purported to support a grading leniency hypothesis but concluded that the research was weak and flawed. In marked contrast, Greenwald (1997 ; Greenwald & Gillmore, 1997a , 1997b) contended that this set of six studies (based on a total of just seven teachers and 12 sections of six classes) provided convincing evidence for a grading leniency effect. These studies presumably would not have met current ethical standards because of the unethical use of deception with no prior consent in natural classroom settings, but Greenwald and Gillmore berated the critics for not repeating the experiments with improved methods and argued that the results were so clear that replications were not needed. In Marsh and Roche's (1997) critique of these studies, the major concerns were (a) flawed designs (e.g., the researchers taught the classes; experimental groups consisted of either students who were randomly assigned within a single class or intact sections of the same class) resulting in limited generalizability and potential researcher expectancy effects; (b) ambiguity in the grading leniency manipulations (e.g., serious violations of students' grade expectations, which normally, regardless of leniency or strictness,are good predictors of actual grade; students' potential awareness or suspicions that inconsistent grading standards had inexplicably been applied by the teacher; presentation and emphasis of the " *un* expected grades" immediately before the collection of SETs-a potentially serious threat to the validity of the SET responses; manipulations that varied the nature of the course or test materials as well as grading leniency), thus undermining a grading leniency interpretation of any observed effects; (c) the typically weak and nonsignificant effects; (d) the disregard

of SET multidimensionality (in four studies, the apparently largest differences were for items specifically about grades and grading fairness-hardly surprising given the nature of the manipulations and, thus, possibly more interpretable as validity than bias); (e) the ethical and substantive generalizability implications of deception research (where the effects may be due to the deception rather than the substantive nature of the manipulation); and (f) evidence of a harshness effect rather than a leniency effect (e.g., Worthington & Wong, 1979 , reported significant differences on 8 of 18 items in a comparison of satisfactory vs. poor grade conditions in support of a harshness effect, but only 1 of 18 in a comparison of good vs. satisfactory + poor conditions. This suggests a nonlinearity whereby students who receive good grades evaluate teaching as low or lower than those who receive satisfactory grades, thus refuting a leniency effect). Although Greenwald (1997) acknowledged that such reservations "deserve serious consideration" (p. 1183), he did not respond to the specific published criticisms, instead characterizing such critiques as speculation. The concerns outlined here (see also Marsh & Roche, 1997) provide a strong basis for treating as mere speculation any conclusion that the effects observed in these studies are due to grading leniency.

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. Groups of students viewed a videotaped lecture, rated teacher effectiveness, and completed an objective exam that was used as the basis for manipulated grade feedback at the start of a second session involving another lecture, SET administration, and exam. The manipulation of grading standards had no effect on achievement and had weak, inconsistent effects on SETs. Whereas these findings do not support a grading leniency effect, the external validity of the grading manipulation in this laboratory study may be questioned. Nevertheless, the design overcomes many of the problems associated with the experimental field studies and offers a promising direction for further research.

d'Apollonia, Lou, and Abrami (1998) reported on a meta-analysis of nine grade manipulation studies, including experimental field studies and laboratory studies. Their meta-analysis was prompted, despite the small number of studies, by the importance of the issue, the controversy surrounding it, and apparently inflated interpretations of these results by Greenwald and Gillmore (1997a , 1997b). d'Apollonia et al. concluded that the average effect size was only .22 (higher ratings associated with higher grades), an effect that they argued was too small to have any practical significance. More significantly, they found an extreme heterogeneity in the effect sizes (which varied from -1.32 to 1.89), again undermining any interpretation of a grading leniency effect. Initial group nonequivalence was the only coded study characteristic that explained much of this heterogeneity (effect sizes were larger when group equivalence was not controlled). d'Apollonia et al. also emphasized that interpretation of the results from just nine studies must be tentative and any recommendation to control SETs for expected grades based on these results is clearly unwarranted.

Path Analysis Studies

Path analytic studies (see Marsh, 1983 , 1987) demonstrate that prior subject interest explains about one third of the grade-SET relation. Because prior subject interest precedes grades, a large component of the grade-SET relation is apparently spurious. This supports a prior characteristics hypothesis.

Howard and Maxwell (1980) conducted an important path analysis of relations between grades, prior student motivation,global SETs, and student self-ratings of their progress on different learning outcome goals (e.g., gaining factual knowledge; learning to apply course materials; developing creative capacities; learning how professionals in the field gain new knowledge) constructed to be appropriate to

all courses. Their analyses were based on an archive of IDEA ratings from a cross-section of different universities. The progress composite was class-average student rating of progress on each goal weighted by teacher importance ratings (using the weights 0, 1, and 2 so that goals that were not important were excluded). In the IDEA system, this composite variable is designed to reflect student growth in learning or achievement in a way that is qualitatively different from the typical student ratings (Cashin & Downey, 1992 ; but see also Marsh, 1995). In the path model, global SETs were posited to be a function of prior motivation, progress ratings, and grades. Replicating Marsh's finding, the modest relation between grades and global SETs (5.3% and 7.3% of variance explained in teacher and course ratings, respectively) was reduced by more than two thirds (to 1.5% and 1.5%) when prior motivation was controlled. Moreover, the grade-SET relation was further reduced (to 0.9% and 0.9%) by also controlling student progress. These results support both the prior characteristics and the validity hypotheses but undermine support for a grading leniency hypothesis.

Gillmore and Greenwald (1994) developed alternative workload and expected grade measures that they related to SET and Perceived Learning measures. SET was an average score based on 11 different rating items (e.g., organization, clarity, participation,assignments, grades, feedback), and Perceived Learning was progress on specific learning goals like those in the IDEA system. Grade items included a typical expected grade item (absolute grades) and a relative grade item that asked students to rate their expected course grade compared with previous course grades. Absolute grades correlated slightly more with SET (.34) and Perceived Learning (.38) than did relative grades (.25 and .32, respectively). Workload was inferred from ratings of intellectual challenge, effort, and involvement, plus two hours items: total hours and valuable hours (number of total hours considered to be valuable). Challenge,involvement, and effort were all positively correlated with SETs (.35, .24, and .14, respectively) and Perceived Learning (.40, .36,.24, respectively). Total hours had nonsignificant positive relations with both outcome variables, but valuable hours was very highly correlated with SETs (.62) and Perceived Learning (.61). Predicting both SET and Perceived Learning from all the workload and grade measures and other background variables (class size, class level, rank, credits), the authors identified three background variables that had significant beta weights for both SETs and Perceived Learning: valuable/total hours (.48 and .44), challenge (.33 and .25), and relative grades (.18 and .27). Involvement (.14) and total hours (-.14) also significantly predicted Perceived Learning. The authors also noted that if relative grades were excluded from the regression equation, absolute grades were significant, and that if challenge was excluded, effort was significant. Whereas valuable hours contributed substantially and positively, bad hours (total hours after controlling valuable hours) were nonsignificant for SET and marginally negative for Perceived Learning. Emphasizing the valuable versus bad hours distinction,they concluded that "overall ratings are predicted by a combination of the ratio of valuable hours to total hours, grades, and the challenge or effort needed to succeed in the course" (Gillmore & Greenwald, 1994 , p. 12).

Gillmore and Greenwald (1994) then fit a structural equation model in which grades (relative and absolute grades) led to Workload (total hours, effort, and challenge) and Overall Evaluation (SET, Perceived Learning, and a global teacher rating, with smaller cross-loadings for challenge and effort). Grades were negatively related to Workload and positively related to Overall Evaluation. Surprisingly, no path was posited between Workload and Overall Evaluation, despite it being strongly implicated by the positive workload effects in Gillmore and Greenwald's multiple regression analyses and by previous research and theory (Marsh & Dunkin, 1992). It is also disappointing that they did not include valuable hours in their path model, particularly given its strong relation to SETs. Nevertheless, Gillmore and Greenwald suggested that the resultant negative grade-workload path may reflect either (a) a simple reality that teachers with more demanding grading standards actually do require more work or (b) that students interpret workload relative to perceived success such that lower expected grades are a sign of a

hard, demanding course (consistent with the interpretation offered by Marsh, 1980 , 1983 , 1987). The authors interpreted the grade relations as a bias but added the important caveat that without controlling "an independent measure of learning we cannot be sure of this conclusion" (Gillmore & Greenwald, 1994 , p. 15). They concluded that overall ratings appear to be influenced by three factors: higher levels of valuable hours, greater levels of challenge, and expected grades, but emphasized that it would be a mistake "to conclude that by giving high grades alone one can assure high ratings" (p. 15). They also speculated about a long-term cycle in which teachers progressively lowered grading standards and workloads in the belief that this would improve SETs but offered no empirical support for these potentially dire implications that require longitudinal comparisons.

Marsh (1982 , 1987) used an alternative path-analysis approach to compare ratings of the same teacher teaching the same course on different occasions. The course offering with higher grades tended to be rated more favorably on each of the SEEQ factors except Workload. Marsh argued that this within-teacher comparison largely controlled grading standards that are typically confounded with student learning and prior characteristics (it is unlikely that grading standards will differ for two offerings of the same course taught by the same teacher, and the mean of expected grades did not differ systematically for the earlier or later versions of these offerings; thus, no tendency toward grade inflation was evident). Although alternative explanations may exist,the results favor a validity hypothesis over a grading leniency hypothesis.

Direct Measures of Grading Leniency

Expected grades are patently not a measure of grading leniency and are used only because no suitable measure of grading leniency is typically available. Surprisingly little research has used alternative direct measures of grading leniency. Marsh and Overall (1979) measured teacher self-perceptions of their own grading leniency (on an *easy/lenient grader* to *hard/strict grader* scale). Leniency relations with both student and teacher evaluations of teaching were small (r s between -.16 and .19) except for ratings of Workload (r s of .26 and .28) and teacher self-ratings of Examinations and grading appropriateness (r = .32). Marsh (1976) also found that self-reported easy graders received somewhat lower overall course and Learning/Value ratings. Hence, results based on direct measures of grading leniency argue against the grading leniency hypothesis.

Summary of Expected Grade Effects

In summary, the small grade-SET relations (about .20 for overall teacher ratings) appear, on the basis of multiple sources of evidence, to be best interpreted from a validity and prior characteristics perspective. Whereas it is possible that a grading leniency effect may produce some bias in SETs, it has been reassuringly difficult to find evidence to support this suggestion,and the size of any such effect is likely to be insubstantial, given the small size of the grade-SET relation and the contribution of other valid factors.

Influences on Grades and Workload: Distinguishing Teacher, Course,and Department Effects

To what extent are workload and expected grades a function of a particular course or department rather than the teacher? Separate course and teacher effects on SETs were estimated with a path analysis on a large database of multiple sets of ratings for the same or different teachers teaching the same or different courses (Marsh, 1987). Within each set, there were ratings of a target course, the same teacher teaching the same course on another occasion, the same teacher teaching another course, a different teacher

teaching the same course, and a different teacher teaching a different course in the same department. Although SEEQ factors (and overall teacher and course ratings) were primarily a function of the teacher (not the course), background variables, including Workload and grades, were substantially a function of the course and department. Teacher and course effects on Workload were approximately equal, and there was also an effect of academic department. Expected grades correlated .36 for different courses taught by different teachers within the same department (a department effect), .48 for the same course taught by different teachers (a course effect), and .40 for different courses taught by the same teacher (a teacher effect). Hence, both Workload and grades are substantially a function of the department and the particular course as well as the teacher who teaches a course.

Study 1: Reanalysis of Greenwald and Gillmore (1997b) Data

Rethinking the Rationale of Greenwald and Gillmore's Thought Experiments

Greenwald and Gillmore (1997a , 1997b) interpreted relations between SETs, expected grades, and workload ratings as support for a grading leniency bias. In Study 1, we critically review these interpretations and reanalyze their published data to more appropriately evaluate their assumptions and interpretations. Greenwald and Gillmore (1997b) based their analyses on a series of simple *thought experiments,* each of which hypothesized varying patterns of relations among SETs, grades, and workload in combination with student achievement and one of four additional causal variables: quality of instruction, student ability, student motivation, and grading leniency. Although provocative, the weakness in this approach is that their conclusions are based on a series of untested, highly implausible assumptions that were not critically evaluated.

Evaluation of the Grading Leniency Model

Greenwald and Gillmore (1997b) hypothesized four different thought experiment models, but they attempted to test only the one model that was fundamentally different from any of those that they posited. In particular, each of their thought models posited student achievement as one of the potential causes of workload, grades, and SETs, but achievement was not even considered in the model they tested. Because grades reflect student achievement and prior student and course characteristics as well as, perhaps,grading leniency, there is absolutely no basis to assume that grades represent only grading leniency. We consider these limitations to be crippling weaknesses in their empirical tests. Nevertheless, given their emphasis on their grading leniency model, we evaluate the internal logic of this model. We contend that their grading leniency model makes problematic assumptions that undermine its credibility:

1. The *work regulation* assumption, which is central to their interpretation of the grade-workload relation, is particularly dubious, overly pessimistic, apparently post hoc, and deceptively loose as an explanatory construct. It implies that (a) students stop working once they achieve their aspired grade such that higher grades lead to lower levels of work and achievement, and (b) low grades motivate students (a no pain, no gain philosophy) to work harder and achieve more whereas high grades have the opposite effect. In contrast, most motivational theories emphasize the superiority of positive feedback and reinforcement in producing effort and persistence (e.g., Covington, 1997 ; Weiner, 1980). Reinforcing students with good grades should increase involvement and effort, not diminish it (indeed, Greenwald and Gillmore reported that involvement was positively related to expected grades). Given this reasonable assumption, if workload has a positive impact on achievement (as the authors assume), then achievement should be positively-not negatively-related to expected grades. Consistent with our contention, Greenwald (1996) previously argued that it "seems reasonable to expect that students should

work harder in courses in which they receive high grades than in ones in which they receive low grades" (p. 8). In contrast to the apparently post hoc work regulation hypothesis, he commented that the negative correlation "between grades and workloads is one finding for which no satisfactory (or at least plausible) theoretical explanation has yet been suggested" (p. 12). The claimed diagnostic value of the model is lost without the dubious work regulation assumption, but we see no theoretical rationale for why their model implies a negative grade-workload correlation. Moreover, the work regulation assumption depends on vague notions of aspiration and the apparent assumption that aspirations do not vary with the course in question (e.g., students may be happy to pass courses perceived as difficult but may be disappointed with a higher grade in courses perceived as easy). In summary, the work regulation rationale underlying their grading leniency model is flawed.

2. Greenwald and Gillmore (1997b) argued that "when expected grade differences are due to grading leniency-and only in this case-there should be a negative correlation between expected grades and course workloads" (p. 744). The whole rationale and interpretation of their thought experiments is predicated on the assumption that grading leniency is the only explanation for the negative grade-workload correlation. Hence, if alternative explanations do exist, then their rationale disintegrates, along with their subsequent bias interpretation of grade-SET relations. This is important because there are many alternative explanations for this negative grade-workload relation that do not involve grading leniency, including ones already discussed in this article and several previously reported by Greenwald and Gillmore: (a) Gillmore and Greenwald (1994) suggested that students interpret workload from the perspective of perceived success such that lower grades are an indication of a hard, demanding course, an explanation similar to Marsh's (1987) suggestion that expectations of receiving low grades logically lead students to conclude that a course is more difficult and demands more work to achieve success. Consistent with this alternative, the most negative grade-workload correlation was for Greenwald and Gillmore's item "the amount of effort needed to succeed in this course"; (b) consistent with their attribution hypothesis, students may attribute lower grades to external characteristics such as course difficulty and heavy workload; (c) the relation could reflect prior student characteristics (e.g.,classes with more able or more motivated students may earn high grades and still not find the course difficult); (d) teachers who are strict graders may actually require more work, making the relation a function of prior course characteristics that have nothing to do with student work regulation processes (as suggested by Greenwald's, 1996 , survey of teachers and similar findings reported by Marsh, 1987). These and other alternative explanations (e.g., d'Apollonia et al., 1998) clearly demonstrate that grading leniency is not the only possible explanation of the negative grade-workload relation. Because Greenwald and Gillmore's (1997b) interpretation of their thought experiments is predicated on this false assumption, the validity of their subsequent conclusions is severely undermined. The failure of any of their thought experiments to test other such explanations-even ones they had proposed previously-undermines this approach.

3. Greenwald and Gillmore (1997b) briefly noted that Marsh (1980 , 1984) also reported a negative grade-workload relation, but they argued that "the full import of the negative relation can become clear only when it is examined in conjunction with evaluative ratings data" (p. 750). In fact, as described earlier, Marsh considered the workload-SET relation in conjunction with multidimensional SET factors and a variety of other background variables and established that (a) much of the grade-SET relation can be explained in terms of prior subject interest (Greenwald & Gillmore posited a motivation effect but did not include this variable in their grading leniency model); (b) workload was positively related to SETs, undermining arguments that students can be seduced into giving favorable SETs by being offered easy courses (Greenwald & Gillmore fixed this path to be zero in their model and excluded their workload items-good hours and involvement-that were most positively related to SETs); and (c) grade-SET

relations vary substantially,depending on the particular SET factor, undermining simplistic bias interpretations and grade satisfaction hypotheses (Greenwald & Gillmore did not consider multiple SET dimensions).

4. We agree with Greenwald and Gillmore (1997b) that their grading leniency model is overly simplistic in ignoring the likely effects of prior student motivation and quality of instruction that must surely influence the relations among SETs,expected grades, and workload. Furthermore, it would have been useful to consider other background variables (e.g., class level,teacher rank, enrollment, coursework mastery) and particularly their measure of valuable hours that were considered by Gillmore and Greenwald (1994) in earlier analyses of these data but that were excluded from their subsequent research.

5. In Greenwald and Gillmore's (1997b) grading leniency model, student achievement is posited to have no direct effect on grades. The grade-achievement correlation, however, is predicted to be negative (grades negatively effect workload, which has a positive effect on achievement, and achievement is negatively related to grading leniency, which has a positive effect on grades). This prediction seems implausible in that achievement and grades should be positively-not negatively-correlated.

6. Particularly problematic is the implication that the SET-achievement correlation is negative (achievement is negatively correlated with grading leniency and grading leniency has a positive effect on SETs that is mediated by grades). The assumption of a negative SET-achievement relation is implausible. Moreover, the multisection validity studies that we reviewed earlier clearly demonstrate a positive SET-achievement relation in settings where grading leniency and prior characteristics are largely controlled. Hence, this prediction, which is based on their grading leniency model, is inconsistent with well-established findings.

As emphasized by Gillmore and Greenwald (1994) , there is no way to separate leniency and achievement effects in grades without including a measure of achievement. Hence, their own prior warning undermines the internal logic of their subsequent study. We suggest that the pattern of results and support for the grading leniency model would have been quite different if achievement had been included in their analyses. Fortunately, their published results provide a potentially useful measure of student mastery that has allowed us to test alternative conclusions: their Perceived Learning measure (student-rated progress on broadly applicable learning goals), which is very similar to the IDEA measure used by Howard and Maxwell (1980) for a similar purpose.

Method

Because Study 1 is a reanalysis of Greenwald and Gillmore's (1997b) results, we refer the reader to that article for details of the sample and procedures. Briefly, they used structural equation modeling to estimate relations among nine variables (see earlier discussion of Gillmore & Greenwald, 1994) posited to represent three latent constructs: (a) SETs, represented by one global teacher rating, a composite teacher-course rating variable (the mean of an idiosyncratic set of 7 specific SET items selected from a pool of 11 so as to maximize internal consistency),and Perceived Learning (student self-ratings of progress on seven learning goals); (b) grades, represented by absolute grades and relative grades (expected grades relative to previous grades); and (c) Workload, represented by hours per week per credit hour, effort needed to succeed, involvement in the class, and challenge presented by the course. Greenwald and Gillmore attempted to fit models positing three latent factors representing all nine variables, but they repeatedly failed to obtain a satisfactory solution with confirmatory and exploratory

factor analyses. Instead, they selected various subsets of variables to represent the three latent factors that resulted in acceptable solutions that constituted the basis of their substantive interpretations. There are, however, serious problems with their strategy that may undermine substantive conclusions:

1. Their analytic approach ignored the well-established multidimensionality of SETs by eliminating the less intercorrelated specific rating items to form an internally homogeneous set of responses, and by treating these multiple items as a single composite. This is problematic in that grade-SET relations vary from close to zero to modestly positive (.30), depending on the specific factor (Marsh, 1987). Although Greenwald and Gillmore (1997a) argued that the original set of 11 items was dominated by a single factor, no tests of unidimensionality were reported, and Gillmore and Greenwald (1994) noted that correlations between some pairs of items were considerably smaller than others (presumably items excluded by Greenwald & Gillmore, 1997b). It is difficult to know what is being measured by such an ill-defined composite (Marsh & Roche, 1997).

2. Consistent with the design of the IDEA program, Howard and Maxwell (1980 , 1982) treated a measure like Greenwald and Gillmore's (1997b) Perceived Learning (self-ratings of progress on specific learning outcomes) as an indicator of student achievement rather than as an SET measure. Greenwald and Gillmore dropped Perceived Learning from their final model because it did not fit with the other SETs but did not consider it as a potential measure of achievement that would have been consistent with their a priori thought experiment model.

3. The elimination of one third of the original variables to get satisfactory results may undermine support for their interpretations. If such post hoc modifications substantially alter substantive interpretations, as may be the case here, then the substantive interpretations should be made cautiously.

Our concerns about the rationale and analyses of models posited by Greenwald and Gillmore (1997b) led us to pursue further analyses of their published results (p. 749, Table 4) using LISREL (Jöreskog & Sörbom, 1993). Our models differed from theirs in that we hypothesized a positive effect of workload on SETs (as previously reported in path models by Marsh, 1980 , 1984 , 1987), whereas they posited no workload-SET path. Like Greenwald and Gillmore, we began with a priori models in which each indicator was allowed to load on only one latent factor, but we then explored post hoc modifications in which additional parameters were freed on the basis of theory and prior research as well as LISREL's modification indices. In contrast to Greenwald and Gillmore's strategy to exclude variables, however, we sought models that fit all their original variables. To facilitate systematic evaluation of the implications of our strategy, we briefly summarize an *audit trail* of the implications of the model modifications. If substantively important differences exist in a priori and a posteriori results, a posteriori solutions should be interpreted with appropriate caution, but if not, a posteriori solutions are defensible (particularly given the large sample size that minimizes capitalizing on chance).

We began with three-factor models with the same latent constructs and causal ordering as posited by Greenwald and Gillmore (1997b) . Following Howard and Maxwell (1980 , 1982) and our earlier discussion, we also posited a four-factor model in which Perceived Learning was treated as a separate factor. On the basis of the validity hypothesis of the grade-SET relation and previous research, Perceived Learning is posited to have a positive effect on grades, Workload, and SETs, and most importantly, the grade-SET path is predicted to be substantially reduced compared with the corresponding path in the three-factor model. These predictions are inconsistent with Greenwald and Gillmore's grading leniency model, which posits no direct effect of student achievement on either grades or SETs and, apparently, no effect of the inclusion of Perceived Learning on the size of the

grade-SET path.

Following Marsh, Balla, and Hau (1996) , we ascertained that solutions were proper, evaluated parameter estimates in relation to predictions, and assessed goodness of fit with the Tucker-Lewis index, the relative noncentrality index, and the chi-square test statistic. The fit of a model is typically concluded to be acceptable if the Tucker-Lewis and relative noncentrality indexes are greater than .9, although these are merely guidelines, and it is valuable to compare the fit for competing models of the same data.

Results

We focus on the comparison of models positing three factors (Workload, grades, and SETs) with models positing four factors (Perceived Learning, Workload, grades, and SETs), and, in particular, on the effects of expected grades and Workload on SETs in these models. Specifically, we posited that Workload would be positively correlated with SETs (opposite to the direction predicted by a workload bias) and that the positive effect of expected grades in the three-factor model would be substantially reduced when the effects of Perceived Learning were controlled in the four-factor model.

Three-Factor Models

The three-factor a priori solution did not fit the data very well (Table 1) and was not proper. Hence, a posteriori solutions were pursued on the basis of LISREL's automatic modification option and an evaluation of the theoretical reasonableness of these added parameters. Although several alternative solutions were considered, it is important to emphasize that the substantive interpretations (based on paths and correlations among the three latent constructs) were similar in each of the models. In particular (see Table 1 and Figure 1), expected grades have a positive effect on SETs (.55) and a negative effect (-.38) on Workload,whereas Workload has a positive effect on SETs (.27). The effects of grades are roughly similar to those reported by Greenwald and Gillmore (1997b , Figure 3). The positive effect of Workload on SETs was not reported by Greenwald and Gillmore, but is consistent with previous research. Several variables loaded on more than one factor (Figure 1 A): Challenge loaded on both the SET and Workload factors (as in Greenwald and Gillmore's final solution), involvement loaded on both Workload and grades (involvement was excluded by Greenwald and Gillmore), and Perceived Learning loaded on all three factors (Perceived Learning was excluded by Greenwald and Gillmore). The reason we found a positive workload-SET effect whereas Greenwald and Gillmore did not is that we used all of their original variables whereas they selectively eliminated variables. At least in this respect, their post hoc modifications fundamentally altered the substantive interpretation of their results, whereas ours did not.

Four-Factor Models

The a priori four-factor solution was fully proper (see Table 1), although the fit was not completely satisfactory. As with the three-factor solution, we used LISREL's modification procedure and theoretical reasonableness to free additional parameters, but these alterations did not alter the substantive interpretations. The two parameters that were added in the final model (see Table 1 and Figure 1 B) were also included in the three-factor model: Challenge loaded on SETs and workload (as in Greenwald and Gillmore's, 1997b, model) and involvement loaded on grade and workload. Substantively, the most important finding from the four-factor model is that the substantial grade-SET path from the three-factor model is completely eliminated (.55 vs. -.07). Also, there are substantial positive paths from Perceived Learning to grades and workload as well as to SETs. Interestingly, workload that is not associated with

Perceived Learning has no positive effect on SETs (the effect is slightly negative, -.10), possibly reflecting the distinction between valuable hours and nonvaluable hours suggested by Gillmore and Greenwald (1994) but not pursued by Greenwald and Gillmore (1997b) . Our new results are consistent with predictions based on the validity hypothesis of the grade-SET relation but contradict the grading leniency model originally posited by Greenwald and Gillmore (1997b) .

Discussion

The two critical findings of Study 1 are that (a) the workload-SET effect is positive in the three-factor model and (b) the positive grade-SET effect in the three-factor model is eliminated when Perceived Learning is controlled in the four-factor model. The first finding invalidates a critical, untested assumption in the Greenwald and Gillmore (1997b) model, and the second contradicts their central conclusion. More specifically, classes expecting higher grades also report better perceived learning outcomes, and this accounts for the higher student ratings (at least in relation to Greenwald and Gillmore's perceived learning measure). This finding is compatible with the validity interpretation of the grade-SET relation but not with Greenwald and Gillmore's grading leniency hypothesis. Importantly, our four-factor model provides a better representation of Greenwald and Gillmore's thought experiment model than does the one that they actually tested, thus undermining the theoretical justification of their empirical tests and providing an alternative explanation of their results.

A critical feature of our disagreement with Greenwald and Gillmore (1997a , 1997b) has to do with the role of student achievement in the interpretation of grade-SET relations. Within their thought experiment models, they seem to accept the need to control prior student and course characteristics and student achievement before attempting to interpret grade-SET relations as grading leniency effects (see also Gillmore & Greenwald, 1994). In particular, they posited student achievement as a causal variable in each of their thought experiments. However, operationalizing a measure of student achievement in this research is a difficult undertaking. This is why multisection validity studies are so important when considering grade-SET relations, as this apparently is the only area in SET research where achievement has been operationalized effectively. Alternatively, we see three broad directions that research can take in addressing this concern:

1. Researchers can "finesse" the issue, arguing that it is not necessary to resolve this problem. We, for example, argue that the grade-SET relation is so small ($r = .20$; and substantially reduced even further by controlling prior student and course characteristics) that the amount of variance left to be explained by grading leniency is very small. Taking a very different approach, Greenwald and Gillmore (1997b) implied that student achievement is unrelated-or even negatively related-to grades, and therefore, it is not important to control for it. As we emphasized earlier, we disagree strongly with their logic.

2. Researchers can experimentally manipulate a theoretically defensible operationalization of grading leniency. Although Greenwald and Gillmore may disagree with previous reviews of grade manipulation studies (e.g., Abrami et al., 1980 ; Marsh, 1987), they probably agree that current ethical standards preclude most such studies. Whereas Dr. Fox-like studies such as the Abrami et al. study may be a viable compromise-certainly one worth pursuing-this may not be a satisfactory option.

3. Researchers can operationalize student learning in path analysis studies and determine how its inclusion affects grade-SET relations. Our strongest area of disagreement with Greenwald and Gillmore is their failure to pursue this approach or to even acknowledge that minimally adequate tests of their thought experiment models required them to do so (as was acknowledged by Gillmore and Greenwald,

1994). Hence, one purpose of our reanalysis was to demonstrate this use of their Perceived Learning measure. Consistent with our predictions and in support of the validity interpretation of the grade-SET relation, controlling Perceived Learning entirely eliminated the positive grade-SET relation. We agree, of course, that a serious concern with our reanalysis is the assumption that student ratings of their progress on a variety of learning outcome goals are adequate representations of achievement. Although Cashin and Downey (1992) provided some justification for this use of progress ratings, we have been critical of the failure of IDEA research to more fully evaluate the construct validity of this measure (Marsh, 1995). Hence, it is important to use appropriate caution in the interpretation of Study 1 results and the related findings by Howard and Maxwell (1980 , 1982) that were based on a similar measure. Obviously, a stronger test would incorporate multiple measures of achievement, at least some of which are not based on student self-reports, to more fully evaluate the construct validity of this measure and our interpretations. This Perceived Learning measure was, however, the best measure of achievement that was available,and it is unjustifiable to claim any support for a grading leniency interpretation of expected grades unless some defensible measure of achievement is included. Although a different measure of achievement might have a smaller impact, the weight of evidence reviewed earlier and our results strongly imply that controlling achievement and prior student and course characteristics substantially reduces grade-SET relations.

The positive workload-SET relation plays an important role in these results. In the three-factor models posited here that are most similar to Greenwald and Gillmore's models, these effects are positive. This is consistent with our previous results and argues against the typical workload bias. Also, this positive workload effect is consistent with two thought experiment models rejected by Greenwald and Gillmore (1997b) , but contradicts their grading leniency model. They excluded this path from their model on the basis of yet another untested a priori assumption, but earlier we outlined theoretical and empirical support for this path based on previous research (Marsh, 1980 , 1983 , 1987). Greenwald and Gillmore (1997b) claimed that unreported models "in which Workload has a direct connection (in either direction) to Evaluation fit much less well with the data" (p. 749) than did models that excluded this path. This claim, however, is curious in that the addition of one more path cannot result in a substantially poorer fit (even if the additional path is zero, the chi-square cannot be any worse). More importantly, inspection of their raw correlations (p. 749, Table 4) reveals that course and instructor ratings were positively correlated to two of their workload measures (challenge, .35; involvement, .25) and almost uncorrelated with the other two (effort, .07; hours, -.12) and that these workload items were even more positively correlated to their Perceived Learning (progress) measure. Furthermore, Gillmore and Greenwald (1994) reported that valuable hours was even more positively correlated with SETs (.61), but this workload item was completely excluded from the Greenwald and Gillmore (1997b) study.

Grades had a negative effect on workload in Greenwald and Gillmore's (1997b) original analysis and our reanalysis. This result is also consistent with Marsh's (1980 , 1983 , 1987) results, as was emphasized by Greenwald and Gillmore. They argued that only a grading leniency model could explain this effect,but earlier we outlined many alternative explanations-including several proposed previously by Gillmore and Greenwald (1994) . Furthermore, their interpretations depend on highly implausible assumptions in their grading leniency model: that rewarding students by giving them higher grades will cause them to work less and to achieve less and that actual achievement is negatively related to grades and SETs. In particular,the predicted negative SET-achievement relation is inconsistent with well-established findings based on multisection validity studies. Although more research is needed to evaluate why expected grades and workload are negatively correlated, the direction of this relation should not be interpreted as support for the grading leniency model.

In summary, Greenwald and Gillmore (1997a , 1997b) argued that relations between grades, workload, and SETs supported a grading leniency interpretation. In essence, their study merely demonstrated that there is a positive grade-SET relation without providing empirical support for their interpretation of this as a grading leniency effect. We criticized their rationale, the theoretical basis of their predictions, and their analyses. Our reanalysis of their data showed that the workload-SET relation was positive and that the negative grade-SET correlation was eliminated when their Perceived Learning (progress on learning outcomes) measure was controlled. More generally, our results are consistent with the validity hypothesis and undermined their claimed support for the grading leniency hypothesis, particularly as operationalized in their thought experiment model.

Study 2: Analysis of SEEQ Data

In Study 2, we use a large database of SEEQ responses for all undergraduate social science courses collected over a 12-year period from a large, private American university to address the following questions:

1. What is the size and direction of grade-SET and workload-SET correlations, and how do they vary as a function of the specific SET factor? A simple grading leniency bias implies that the relations will be similar for different SEEQ factors,whereas at least some patterns of differences are more consistent with a validity or, perhaps, a prior characteristics hypothesis.

2. Are grade-SET and workload-SET relations strictly linear? Whereas Marsh (1987) specifically demonstrated that the SET-enrollment relation is nonlinear, we know of no previously published research specifically evaluating the nonlinearity of relations involving expected grades, workload, and other background characteristics. Following proposals by Marsh (1987) , theoretical accounts by Marsh and Dunkin (1992) , and suggestions by McKeachie (1997a , 1997b),we hypothesized that the positive workload-SET relation evident over most of the workload range would level off or even decline for extremely high levels of workload. Although we know of no empirical tests of the nonlinearity of grade-SET relations,there is some theoretical and empirical basis for this prediction. Our interpretation of the attribution theory suggests that students externalize low grades by attributing them to external causes but internalize high grades by attributing them to internal causes. Also, some of the grade manipulation studies reviewed earlier might implicate a grading harshness effect instead of a grading leniency effect. Finally, McKeachie (1997b) speculated that excessively high grades might lead to lower SETs. This is also a substantively interesting question because a grading leniency bias implies at least a linear effect and may imply that the grade-SET function should be steepest when grades are highest (i.e., the way to get high ratings is to give high grades). In contrast, the attribution hypothesis implies that the function is steepest for low grades (because students attribute these to external factors, such as poor teaching as well as, perhaps, course difficulty and poor exams) and flatter for high grades (because students attribute these to internal factors, such as their ability,effort, and good study strategies).

3. What are the implications of controlling expected grades for prior grade point average (GPA)? Greenwald and Gillmore (1997b) emphasized the differences between the typical absolute grades and a relative measure of grades based on ratings of whether expected grades are higher or lower than previously received grades. They interpreted the relative grade measure as a better indicator of grading leniency than absolute grade, but it is also reasonable to interpret this as a measure of better learning in that students perform better than they have in the past. Also, relative grades are a curious mixture of grades and a prior student characteristic (prior GPA). A possibly surprising implication of their interpretation is that prior GPA has a negative effect on SETs (i.e., relative grades = grades - prior GPA

is more highly correlated to SETs than grades alone, so the effect of prior GPA on SETs must be negative). Here, we explore the distinction between relative and absolute grades by incorporating both grades and prior GPA into the same path models.

4. How do mean levels of expected grades, workload, and SETs and relations among these variables vary over an extended period of time? Greenwald (1998) speculated that grade-SET relations may have changed over the past few decades. Extrapolating from the grading leniency model, Greenwald (1996) speculated that

> if students tend to choose courses taught by reputedly lenient instructors, then there can be an erosion of the difficulty level of courses as students oversubscribe high-grading, easy courses relative to lower-graded, more difficult courses. This would be an educational analog of Gresham's Law in economics (counterfeit currency drives genuine currency out of circulation). Further,students will likely respond to strict instructors with low ratings,which can put pressure on those instructors to shift toward greater leniency. (p. 14)

Also, if teachers become more differentiated in grading leniency over time, then the grade-SET relation may increase. Whereas we disagree with implications of this doom and gloom scenario, the longitudinal data considered here provide a uniquely appropriate test of these predictions based on their grading leniency model.

Method Procedures

During the period 1976-1988, SEEQ was routinely administered in all 10 academic departments constituting the Division of Social Sciences at a large private university in the western United States. SEEQ was required and was an important basis of personnel decisions. This unique database allowed us to evaluate grades, workload, SETs, and relations among these variables over an extended period of time, starting from when SEEQ was first introduced. SEEQ forms were typically distributed to staff members shortly before the end of each academic term, administered and collected by a student in the class or by a member of the administrative support staff according to printed instructions, and taken to a central office where they were processed. This program,the SEEQ instrument on which it is based, and research that led to its development were described by Marsh (1987 ; Marsh & Dunkin, 1992 ; Marsh & Roche, 1994 , 1997). Classes included 5,433 undergraduate classes offered in 10 departments of the Division of Social Sciences, which were taught by regular academic staff. Excluded were graduate level courses, courses taught by teaching assistants, and courses that were evaluated by less than six students.

Materials

SETs are summarized by the nine SEEQ factors (Learning/Value,Instructor Enthusiasm, Organization Clarity, Group Interaction,Individual Rapport, Breadth of Coverage, Examinations/Grading,Assignments/Readings, Workload) and the two overall summary (overall teacher and overall course) ratings (see Marsh, 1987 , for the wording of SEEQ items) that have been supported by more than 40 factor analyses (e.g., Marsh, 1983 , 1984 , 1987 ; Marsh & Hocevar, 1991 ; Marsh & Roche, 1994). Factor scores represent the SEEQ factors, whereas the overall teacher and course ratings are based on responses to separate individual items. For present purposes, the Workload factor is considered to be a background variable as well as a SEEQ factor. Other background variables are prior GPA, 1 (*below 2.5*) to 5 (*above 3.7*); year in school (college), 1 (*freshman*) to 5 (*graduate*); prior subject interest (level of interest in this subject prior to this course), 1 (*very low*) to 5 (*very high*);

enrollment; and expected grades, 0 (*F*) to 4 (*A*). Items used to infer workload were course difficulty, 1 (*very easy*) to 5 (*very hard*); course workload, 1 (*very light*) to 5 (*very heavy*); course pace, 1 (*too slow*) to 5 (*too fast*); and hours per week required outside of class, 1 (*0-2*) to 5 (*12+*). All analyses were conducted on class-average responses.

Analyses

Initial analyses examined linear and nonlinear relations between the 11 SEEQ scores and six background variables. To facilitate interpretations, all SETs and background variables were standardized ($M = 0$, $SD = 1$) and quadratic and cubic components of each background variable were determined by squaring and cubing *z* scores representing each background variable (see Aiken & West, 1991). Adopting a hierarchical approach, variance attributable to the linear component was partialed out of the quadratic and cubic components, whereas variance attributable to the quadratic component was partialed out of the cubic component so that each component was mutually uncorrelated with the other components.

Path analyses were conducted relating the five background variables to overall teacher ratings with the SPSS version of LISREL 7 according to an a priori path model. Because all constructs were represented by a single variable and all possible paths were hypothesized, the solution had $df = 0$ and was necessarily able to fit the data. For present purposes, relations between the variables in this model were summarized as direct effects (path coefficients),indirect effects (effects mediated through intervening variables),and total effects (the sum of direct and indirect effects).

Results

Consistent with our emphasis on the multidimensionality of SETs,we begin by evaluating relations between grades, workload, each of the SEEQ factors, and other background variables. Initially, we look at simple linear and nonlinear relations for each of the background variables (Table 2 and Figure 2), and then we examine the effects of grades and workload in combination with other background variables (Tables 3 and 4). Finally, we combine these analyses in a structural equation model of relations between background variables and overall teacher ratings (Table 5 , Figure 3) that have been the focus of most previous research. Interpretations of results are based in part on the path model (Figure 3) in which prior GPA, year in school, prior subject interest, and enrollment are considered to be prior characteristics that precede grades, Workload, and SETs. Following the logic of Marsh (1980 , 1983), Greenwald and Gillmore (1997b) , and Study 1 of this investigation, expected grades are posited to precede workload, which are followed by SETs.

Class-Average Grade Expectation (Grade) Relations

There are systematic differences in the sizes of relations between expected grades and the various SEEQ factors (Table 2 ; see also Tables 3 and 4). Grades are modestly but most highly correlated with Learning, Exams, and Group Interaction (*r* s ≈ .30) but nearly uncorrelated with Organization, Enthusiasm, and Breadth of Coverage (*r* s < .10). Grades are only modestly correlated with overall teacher ratings ($r = .198$), which have been the focus of most research, and slightly more highly correlated with overall course ratings ($r = .25$). This pattern of results argues against a simple bias hypothesis in that the relations differ so much depending on the SEEQ factor. Also, the pattern of differences is inconsistent with a bias hypothesis. Thus, for example, if grades were operating as a bias factor that rewards teachers for giving higher grades, overall teacher and Enthusiasm ratings should be more highly correlated with grades than the Overall Course and Learning ratings, but the results oppose

this pattern. Instead, Learning is most highly correlated with grades (classes with higher grade expectations report more valuable and challenging learning experiences), and this supports a validity hypothesis.

Nonlinear relations.

The polynomial regressions, particularly those based on the overall teacher ratings emphasized in SET research, demonstrate significant nonlinearity in grade-SET relations (see Table 2). Inspection of Figure 2 A indicates that there is a small (inverted U) quadratic component in the relation whereby the grade-SET function is most positive at the lower end of the grade continuum but is relatively flat over much of the grade continuum and even has an inflection point near the top of the grade continuum (at 3.81; 2.2 *SD* above the mean GPA). In a separate analysis based on classes with average or above-average grades,grades are almost uncorrelated with overall teacher ratings ($r = .06$). This finding is substantively interesting, in part because we are unaware of the nonlinearity in the grade-SET relation being the focus of previous research. More importantly, the finding has momentous implications for interpretations of the grade-SET relation. In particular, in contrast to implications in the term *grading leniency,* SETs are not systematically related to grades when grades are above average. Instead, it is only when grades are well below average that the grade-SET function is relatively steeper. Hence, these results provide particularly strong evidence against the claim that students reward teachers for giving them exceptionally high grades. The results are, however, consistent with an attribution hypothesis that predicts that students will attribute good grades to internal causes and poor grades to external causes (including poor teaching).

Expected grade relations after controlling other background variables.

Consistent with previous research, there is a modest negative relation between grades and workload ($r = -.29$). In contrast, grades have modest positive relations (Table 5) with prior GPA (.33),year in school (.27), and prior subject interest (.24). Thus,expected grades are higher when the class-average GPA is higher,when the class-average year in school is higher, and when the class-average prior subject interest is higher. These correlations between grades and other background variables demonstrate that grade relations cannot be appropriately evaluated without controlling these prior characteristics of students and the course and are consistent with earlier findings that expected grades are more a function of the particular course and discipline than of the particular teacher who is teaching a course.

In Table 3 , we examine the proportion of the variance in SETs associated with expected grades (grade variance components) that can be explained by variables that logically precede it (see Figure 3). Because so much variance is explained by prior subject interest (Marsh, 1987), we consider it separately and in combination with GPA, year in school,and enrollment (linear and quadratic components). Of particular interest are the three SEEQ factors that are most highly correlated with expected grades: Learning, Group Interaction, and Exams.

For Group interaction, expected grades explain 9.2% ($r^2 = .303^2$) of the variance, but this grade variance component drops by more than half (from 9.20 to 4.50) when GPA, enrollment, and year in school are controlled. Not surprisingly, more advanced classes (i.e., classes in which mean year in school is higher) and smaller classes tend to have higher Group Interaction ratings and also tend to have higher grades. The additional control for prior subject interest further reduces the grade variance component to 3.65. Thus, the majority of the grade-group interaction relation can be explained in terms of the other background variables, particularly year in school and enrollment.

For Learning ratings, the grade variance component is reduced by more than one third (9.93 to 6.03) by controlling GPA, enrollment,and year in school. The additional control for prior subject interest produces a further substantial reduction (to 3.52). Controlling prior subject interest alone reduces the grade effect by 62% (from 9.93 to 3.78). Hence, as with Group Interaction ratings, the majority of the grade effect on Learning is explained in terms of preceding variables, but most of the reduction for Group Interaction is due to year in school and enrollment, whereas most of the reduction for Learning is due to prior subject interest.

For the SEEQ Exam factor, the grade variance component is reduced somewhat (9.83 to 7.61) by controlling GPA, enrollment, and year in school, and reduced somewhat further by also controlling prior subject interest (to 7.00). However, in comparison to Learning and Group Interaction, these variables explain a much smaller portion (29%) of the grade variance component. This finding that the grade-exam effect is the largest effect after controlling prior variables may also be consistent with attribution theory (if I get lousy grades the test must be bad, inappropriate, or unfair-whether or not these attributions are true). An interesting speculation based on these results is that the negative effect of grades on overall teacher ratings may be mediated by students' ratings of the quality of examinations and grading. To evaluate this speculation, we included Exams (defined by three items: Feedback on exams was valuable, evaluations of student work were fair and appropriate, and examinations and graded materials tested course content as emphasized by the teacher) in models that contained grades and overall teacher ratings. When only grades and Exams were related to overall teacher ratings, the grade effect on overall teacher ratings was nonsignificant. When all background variables and Exams were included in the model, the grade-overall teacher path was slightly negative. The interpretation of these results depends in part on underlying assumptions of the causal model. If, for example, grades are assumed to precede student perceptions of the Exams and Exams are one of the components that students use in determining their overall teacher ratings, then positive grade effects on overall teacher ratings are mediated by Exams. Although other interpretations may be plausible,the results offer further support for an attribution hypothesis and demonstrate why SETs cannot be adequately understood if their multidimensionality is ignored.

Expected grade relations in overall teacher ratings.

Not surprisingly, the grade variance component for overall teacher ratings is smaller than those for SEEQ factors with the largest variance components (Learning, Group Interaction, and Exams), but larger than those for Organization, Breadth of Coverage,and Enthusiasm which are almost unrelated to grades. Whereas the grade-overall teacher relation appears to reflect an average relation, it cannot adequately reflect the diversity of grade-SET relations for specific SEEQ factors. This argues for the use of the specific factors in addition to the overall rating,particularly if there is a desire to understand the nature of grade relations. However, because much previous research has focused almost exclusively on overall teacher ratings, and because they are sometimes the sole basis for representing SETs for some applications, overall teacher ratings deserve special attention.

Expected grades alone explains a modest 3.9% ($r^2 = .198^2$) of the variance in overall teacher ratings. This variance component is reduced slightly by controlling GPA, year in school, or enrollment individually (Table 3) or in combination, but is reduced by 45% (from 3.9 to 2.2) by controlling the effects of prior subject interest. Because prior subject interest comes before grades in the path model (Figure 3), the grade relations that can be explained by prior subject interest are spurious effects of grades. Students in classes in which prior subject interest is higher tend to expect higher grades and to

rate teaching as more effective.

Although not the focus of previous research, the combined effects of GPA and expected grades are of particular interest in relation to claims by Greenwald and Gillmore (1997b) . They examined what they referred to as absolute and relative grades. Their absolute measure is like the typical expected grade measure, but their relative measure asks students whether they expected to receive a higher or lower grade than they had received in the past. Because grades are typically higher in upper-division courses than in lower-division courses, grades in current courses are likely to be higher than grades in previous courses. The question is whether prior GPA contributes to the prediction of SETs beyond the contribution of grades. If controlling prior GPA reduced the grade-SET relation, then this portion of the grade-SET relation is a spurious effect and should be discounted. Greenwald and Gillmore, however, implied a suppression effect in that grades controlled for prior GPA (their relative measure) is actually more strongly related to SETs than grades alone (their absolute measure).

Although previous SEEQ research has not considered relative grades, the simultaneous effects of prior GPA and grades have been considered. However, because the direct effects of prior GPA have consistently been very small in large studies involving many background variables (including grades), these relations have not received much attention. Greenwald and Gillmore's relative item (the difference between current and previous grades) is conceptually similar to the effects of grades after controlling the effects of prior GPA. (Gillmore & Greenwald, 1994 , originally considered a relative measure, defined as the difference between grades and class-average prior GPA, that was based on actual student records but abandoned it because of difficulties in obtaining the information.) According to the logic of the Greenwald and Gillmore measure, controlling the effects of prior GPA should result in an increased beta weight associated with grades-a suppression effect. Our results, however, provide little support for these predictions in that the beta weight associated with grades (.20; see Table 4) does not change when GPA is added to the prediction equation. Hence, the modest GPA effects are largely mediated by grades; higher expected grades tend to occur in those courses where students have higher prior GPAs. These results provide no support for the Greenwald and Gillmore interpretations of relative grades, although it would be interesting for them to pursue similar analyses with their data based on actual prior GPAs to determine whether their own data support their interpretation.

Workload Relations

Consistent with an emphasis on the multidimensionality of SETs,there are systematic differences in the sizes of relations between workload and various SEEQ factors (see Table 2). Workload is modestly but most positively related to Assignments ($r = .26$) and Learning (.17) but is not significantly related to Group Interaction and Individual Rapport. Workload is also modestly correlated with overall teacher (.19) and overall course (.25) ratings. Although the pattern of relations is intuitively reasonable, the particularly important feature of these relations is that they are positive, not negative. Courses perceived to be more difficult, to have a heavier workload,to require more work outside of class, and to move at a faster pace tend to receive more positive ratings. Hence, the direction of this relation is precisely opposite to that posited by a typical bias hypothesis (i.e., that teachers are rewarded with higher ratings for making fewer demands on students).

Nonlinear relations.

There is also a modest (inverted U) quadratic component to the workload-SET relation. For overall teacher ratings (see Table 2 and Figure 2 B), the function is positive for most of the workload

continuum, but the function levels off and has an inflection point in the upper half of the workload continuum (at 1.15 *SD* s above the mean workload). The nature of this nonlinearity is reasonably consistent across the SEEQ factors, but there are some differences. Thus, for example, the nonlinear component of the workload-learning relation is not significant such that Learning increases linearly with increasing workload. Workload is nearly unrelated to Group Interaction, but the small quadratic component is positive rather than negative. However, for the overall course rating and six of the eight specific SEEQ factors, the nature of the workload-SET function is like that shown in Figure 2 B for the overall teacher ratings (i.e., SETs increase over most of the workload continuum, level off, and decline slightly for very high workload levels). Although not previously tested (to our knowledge),the nature of this nonlinearity is consistent with a priori predictions based on Marsh and Dunkin's (1992) theoretical account of the relevance of workload to effective teaching as well as some discussion by other researchers reviewed earlier.

Variance explained by workload after controlling other background variables.

Workload has small positive relations with prior GPA (.13; see Table 5) and prior subject interest (.19) but is almost unrelated to year in school and enrollment. Workload tends to be greater for classes in which students have higher GPAs and for classes in which prior subject interest is greater. Workload is also negatively related to grades in that students expecting lower grades perceive the course as being more difficult. Because of these correlations between grades and other background variables, workload relations need to be evaluated in combination with these other prior characteristics of students and the course.

As with expected grades, we examine how much variance explained by workload can be explained by variables that logically precede it (see Figure 3). The largest workload variance component is for Assignments (see Table 3). Controlling GPA, year in school, and enrollment reduces this relation only modestly (6.66 to 5.98), but controlling prior subject interest reduces the relation more (6.66 to 4.79). However, even controlling all these background variables reduces the workload variance component by only 30% (6.66 to 4.66). Hence, quality of Assignment ratings are higher in classes with heavier workloads, and less than one third of this relation can be explained by other background variables considered here. This is consistent with observations that Assignments make up a substantial proportion of the work that is done outside of class and with the Gillmore and Greenwald (1994) finding that the majority of work done by students is perceived as valuable work. Again, this seems to support a validity hypothesis, not a bias hypothesis.

The next largest workload variance component is for the overall course rating. Again, controlling GPA, year in school, and enrollment reduces this variance component modestly (6.35 to 5.36),but controlling prior subject interest reduces it more substantially (6.35 to 3.41). Courses with higher levels of workload receive higher overall course ratings, but nearly half of this relation can be explained in terms of background variables, primarily prior subject interest.

Workload effects in overall teacher ratings.

The workload variance component is larger for the overall teacher rating than for all but one of the specific SEEQ factors (Assignments). Although the workload variance component for overall teacher ratings is modest (3.72), it is reduced by 41% (from 3.72 to 2.21) by controlling GPA, year in school, enrollment, and particularly prior subject interest (Table 3). Controlling the effects of GPA and enrollment reduces the variance component associated with workload only slightly, but controlling prior

subject interest reduces it by more than one third. Because prior subject interest comes before workload in the path model, this proportion of the workload effect is spurious. It is, however, important to reiterate that these workload effects-even after controlling the effects of prior subject interest-are positive; teachers teaching courses with heavier workloads tend to be evaluated as more effective. Hence, the direction of this effect is opposite to that typically posited as a potential bias to SETs.

Of particular interest is the complicated pattern of effects attributable to the combination of workload and expected grades (Table 4). Controlling the effects of grades increases the direct effects of workload (the beta weight increases from .19 to .28), and controlling the effects of workload increases the direct effect of grades (the beta weight increases from .20 to .28). This is a particularly dramatic example of the unusual occurrence of mutual suppression (see discussion by Cohen & Cohen, 1983). This occurs because both workload and grades are positively related to SETs but negatively related to each other. It also complicates interpretations of their effects. Because workload follows grades in the path model, the substantially increased direct effect of workload is the same as the total effects of workload. For expected grades, however, the total effects are substantially less than the direct effects. That is, the total effects of grades are composed of a moderate positive direct effect and a counterbalancing negative indirect effect that is mediated by workload. Hence, the total positive effects of workload are substantially greater than the total positive effects of grades even though the direct effects of these two variables are similar.

Relations With Other Background Variables

We examine several other background variables that are sometimes considered to be potential biases, that are correlated with workload and grades, and that logically precede grades and workload.

Enrollment.

Enrollment, not surprisingly, is most negatively correlated with Group Interaction (-.33) and Individual Rapport (-.23) but is not significantly related to Breadth of Coverage, Enthusiasm,and Organization (see Table 2). Enrollment (see Table 5) is negatively related to year in school and, to lesser extents, prior GPA and prior subject interest but is nearly uncorrelated with grades and workload (see Table 2). There is a modest (U-shaped) quadratic component for overall teacher ratings (see Figure 2 C) such that SETs initially decline over the lower range of enrollments. There is,however, an inflection point such that overall teacher ratings begin to increase for large enrollments, and classes with the largest enrollments tend to be rated as high as or higher than small classes. The high ratings for very large classes, however, are based on relatively few data points and may reflect the drawing power of a few "star" teachers (i.e., teacher reputations induce higher enrollment rather than enrollment causing the ratings). More generally, once enrollments reach some critical value where small-class techniques are no longer appropriate, teachers may adopt appropriate large-class techniques that improve the quality of instruction, possibly explaining the nonlinear relations.

Prior subject interest.

Prior subject interest is more highly correlated with SETs than are any of the background variables considered here (see Table 2). In particular, the correlation with Learning (.53) is very large compared with the next highest correlations, for Group Interaction (.28) and Assignments (.23). Prior subject interest is more highly correlated with both overall ratings than are other background variables, although the correlation with overall course rating (.38) is higher than the correlation with the overall teacher

rating (.23). Prior subject interest (see Table 5) is positively correlated with prior GPA, year in school, workload,and grades but negatively correlated with enrollment. There is little nonlinearity in prior subject interest relations,particularly for overall teacher ratings (see Table 2).

Prior GPA.

GPA is modestly but most positively correlated with Learning (.17) and Group Interaction (.15) but is slightly negatively correlated with Organization (-.05) and not significantly related to Breadth and Individual Rapport (see Table 2). Prior GPA is positively correlated with year in school (.30; see Table 5), prior subject interest (.28), grades (.33), and, to a lesser extent, workload (.13), but is negatively correlated with enrollment (-.15). For the overall rating items in particular,there is little nonlinearity in relations with GPA. Almost all of the GPA relation is mediated by grades: Classes in which students have higher (prior) GPAs have high grades and grades are positively related to SETs, particularly the Learning factor but also, to a much smaller extent, overall teacher ratings.

Class-average year in school.

Class-average year in school is modestly correlated with Group Interaction (.30) and, to lesser extents, Learning (.18) and Exams (.14) but is not significantly related to Organization, Enthusiasm,and Breadth (see Table 2). Year in school is positively correlated with prior GPA (.30), prior subject interest (.25), and grades (.27), but the largest correlation is the negative relation between year in school and enrollment (-.47). For the overall rating items in particular, there is little nonlinearity in the relations.

Structural Equation Model

Because the various background variables are moderately intercorrelated, it is important to examine a structural equation model in which relations for all the background variables are considered simultaneously (see Figure 3). Because of the nature of the effects (particularly the suppression effects), it is important to examine indirect (mediated) and total effects as well as the direct (unmediated) effects. Effects are presented in Figure 3 whenever total or direct effects are greater than .10 (all effects are presented in Table 5). Prior characteristics (prior GPA, year in school,prior subject interest, and enrollment) have modest effects on grades and workload. Two of these prior characteristics-prior subject interest and GPA-have direct or total effects on overall teacher ratings of .10 or greater. The total effect of prior subject interest on the overall teacher rating is .22, but some of this effect is mediated through grades and workload (see Table 5). The total effect of GPA (-.02) is negligible, but this total effect represents a negative direct effect (-.10) and a positive indirect effect (.08) that is mediated through grades and, to a lesser extent,workload (see Table 5). As noted earlier, this suppression effect is negligible when only GPA and grades are in the model or when only GPA and workload are included. It is only when both grades and workload are in the model that there is a modest negative direct effect of prior GPA on SETs and a corresponding positive indirect effect. The inclusion of the other prior characteristics has little effect on this general pattern of results.

The total effect of expected grades on overall teacher ratings is .17, consisting of a positive direct effect (.28) and a negative indirect effect that is mediated through workload (-.11). The positive direct effect of grades is also qualified in that there is some nonlinearity in the grade relation such that the grade-overall teacher function is relatively flat in the top half of the grade distribution.

Workload has a positive direct effect on overall teacher ratings (.25). Because there are no variables

between workload and SETs, the direct and total effects are the same. This positive workload effect is, however, qualified to some extent by the significant quadratic component.

In summary, on the basis of the total effects, higher overall teacher ratings are associated with higher workload (.25), higher prior subject interest (.22), and higher expected grades (.17).

How Expected Grades, Workload, Overall Teacher Ratings, and Their Relations Vary Over Time

A unique aspect of this study is that the data are based on the first 12 years that SEEQ was used at this institution. This is important for evaluating interpretations of the potential biases that are allegedly due to expected grades and workload and, perhaps even more important, to long-term implications of these interpretations. If teachers are motivated to try to manipulate SETs by reducing workload, then there should be a steady decline in the workload levels over time. If teachers try to manipulate SETs by using more lenient grading standards, then there should be systematic increases in grades. Greenwald and Gillmore (1997b) also argued that more lenient grading standards will lead to reduced student workload. To the extent that some teachers use these strategies and that the assumed biases actually do exist, the grade-SET relations may become more positive over time and the workload-SET relations may become more negative over time. Although both strategies would constitute a serious threat to the validity of SETs and decisions based on them, the concerns about workload (Greenwald & Gillmore, 1997b) are perhaps even more serious than the grading leniency effect that has received so much attention. If either of these dubious strategies is successful, then it also follows that SETs should increase over time. Although we know of no previous research designed to evaluate these doom and gloom implications of grading leniency and workload biases, Greenwald and Gillmore (1997b) emphasized the importance of such research. Our analyses are uniquely relevant to pursuing these predictions about these changes over time.

Overall teacher ratings increase slightly over time (Table 6), but the linear effect of time explains only 0.25% of the variance, and the nonlinear components are not significant. Perceived workload increases-not decreases-over time, although the linear effect of time is modest (1.29% of the variance) and the nonlinear components are not significant. There is no linear effect of time in the expected grades, although there is a small quadratic effect (0.35% of the variance) in which grades increase slightly and then fall again.

Correlations between overall teacher ratings, expected grades,and workload were also computed separately for each of the 12 years. To determine, for example, whether grade-overall teacher relations vary over time, grades, time (linear, quadratic, and cubic components), and Time × Grade interactions were included in a regression equation predicting the overall teacher rating. The linear, quadratic, and cubic Time × Grade interaction components are small and nonsignificant; there are no changes over time in the grade-overall teacher relations. Similarly, there are no systematic differences over time in relations between overall teacher ratings and workload or between workload and expected grades.

In summary, SEEQ data from over a decade at the same institution provide absolutely no support whatsoever for doom and gloom implications derived from grading leniency and workload bias theories. Workload increased slightly over this period; it did not decrease as was implied by strategies based on a workload bias. Grades neither systematically increased nor decreased over this time period. Also, correlations between grades, workload, and overall teacher ratings were stable over time.

Discussion

Popular myth implies that teachers can manipulate students into giving them favorable ratings by offering less demanding courses and grading more leniently. Study 2 addressed a series of questions to facilitate interpretations of workload-SET and grade-SET relations. Workload-SET relations, as reported in other research, are positive, not negative. The positive direction of this relation argues against a workload bias whereby teachers are rewarded with higher SETs when they offer easier, less demanding classes. The positive direction of the workload effect makes a workload bias untenable.

Grade-SET correlations are modest, varying from 0 to .30 for different SEEQ factors. Consistent with previous research, the grade-overall teacher correlation is about .20. The modest size and the systematic variation of these correlations across different SEEQ factors argue against a simple grading leniency bias. The highest grade-SET 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. Grades are moderately correlated with Group Interaction, but this relation is reduced substantially by controlling year in school and enrollment. This also argues for a prior characteristics hypothesis. However, the moderate correlation between grades and Exams (student ratings of feedback value,fairness and appropriateness of evaluation methods, and testing the content as emphasized) is reduced by less than 30% by controlling background variables considered such that after controlling other background variables, the grade-exam relation is clearly the largest effect of grades. Hence, for exams,a grading leniency bias, or perhaps a self-serving attribution effect (whereby students blame their poorer grades on the quality of examinations) may be viable. Because student achievement has not been controlled, however, this relation may be a valid effect in that teachers who give low grades may also give poor exams. If there were a grading leniency effect such that teachers are unduly rewarded for giving undeservedly high grades, one might expect overall teacher and teacher Enthusiasm-rather than overall course and learning ratings-to be more highly correlated with grades, but the pattern of differences was exactly opposite to these expectations. One particularly important, new piece of information undermining support for a grading leniency bias interpretation is the relatively large nonlinear component of the grade relation whereby grades are nearly uncorrelated with overall teacher ratings for classes receiving grades at or above the mean grade. Finally,the lack of changes in expected grades over more than a decade argues against long-term implications based on a grading leniency effect.

With the possible exception of the SET-enrollment relation,researchers have not systematically explored the nonlinear relations between SETs and different background variables. The nonlinearity of grade-SET and workload-SET relations, however, has important implications. On the basis of Marsh and Dunkin's (1992) theoretical description of how workload should be related to effective teaching, it was predicted that the relation should be nonlinear; courses that are too easy do not adequately challenge students, but extremely overloaded students are not effective learners. Although these predictions have not been previously tested, our results support them in that the workload-SET relation showed a positive linear effect and a small (inverted U) quadratic component. Speculation based on attribution theory and some findings from the manipulated grade studies suggest that there may be some nonlinearity in grade-SET relations in which a grading harshness effect may be stronger than a grading leniency effect. It was noted that the grading leniency hypothesis apparently implies that classes getting the highest grades should be particularly likely to reward teachers with higher SETs. Whereas there was a modest nonlinear component in the grade-SET relation, the nature of this effect was opposite that predicted by a grading leniency hypothesis. The grade-SET function was most steeply positive for the lowest grades, was relatively flat for grades that were average or above, and actually showed an inflection point such that the slope was slightly negative for the highest grades. Hence, the nature of the nonlinearity in the

grade-SET function is inconsistent with a grading leniency bias.

A unique aspect of this investigation is that data consisted of responses from the first 12 years during which SEEQ was used in this setting. If teachers tried to manipulate SETs in relation to grading leniency and workload biases (whether or not they actually exist),it might be expected that grades should systematically increase,that workload would systematically decrease, and, perhaps, that grade-SET relations would increase over time. Evaluation of the longitudinal results, however, demonstrated that SETs, grades,workload, and all relations among these variables were stable over time.

Summary and Discussion

The results of Study 2 and our review of previous research show that there is a consistently small positive correlation of about .20 between global SETs and expected grades. There are at least three competing interpretations of this relation that have very different implications: a validity hypothesis, a prior characteristics hypothesis, and a grading leniency hypothesis. Emphasizing the multidimensionality of SETs and a construct validity approach to evaluate this grade-SET relation, we found clear support for the validity and the prior characteristics hypotheses but limited support for the grading leniency hypothesis. A number of features of this literature argue that any potential effect of grading leniency must necessarily be very small. Importantly, the grade-SET relation itself is very small. There is clear evidence that most of this small relation can be explained in terms of the validity and the prior characteristics hypotheses. Hence, at most, grading leniency is able to explain only a small portion of a very small grade-SET relation.

Study 1 was prompted by the Greenwald and Gillmore (1997a , 1997b) studies claiming to show a grading leniency bias. A critical evaluation of the logic underlying their arguments showed that they were based on unreasonable and largely untested assumptions. Thus, for example, implicit or explicit in their account are the assumptions that receiving good grades causes students to work less and achieve less, that good teachers should give students particularly low grades at the start of a course in order to motivate them (a no pain, no gain philosophy), and that achievement is negatively related to grades and SETs. Importantly,there were critical differences between their a priori grading leniency model and the model they actually tested. In particular,they proposed that grading leniency-poorly operationalized as expected grades-has a positive effect on SETs that is independent of student achievement, and their model assumed that prior characteristics (e.g., prior subject interest) had no effect on these relations. In their empirically tested model, however, they did not include a measure of achievement, nor did they control any prior student or course characteristics. Their model merely demonstrated that grades and SETs are positively correlated. They provided no tests of their alternative thought experiment models,nor did they evaluate the effects of grades on SETs after controlling the effects of achievement or other prior characteristics, such as prior subject interest. This limitation of their study is important because they had a perceived learning measure (students' self-ratings of progress on learning outcomes) like the measure previously used by Howard and Maxwell (1980) . Consistent with the original Howard and Maxwell results, our reanalysis of the Greenwald and Gillmore data shows that controlling for this measure of perceived learning largely eliminates the grade-SET relation. This finding is consistent with the validity hypothesis and undermines support for a grading leniency hypothesis in that the higher grades associated with higher SETs may actually reflect student learning.

A particularly problematic aspect of the Greenwald and Gillmore (1997b) study is their failure to include workload variables that they had previously reported to be substantially and positively related to SETs. Gillmore and Greenwald (1994) offered the intriguing and potentially valuable distinction between

"valuable" hours and "bad" hours. Despite the central implications of this distinction to understanding workload and its relation to SETs, they ignored this distinction in their subsequent research. The failure to retain this distinction apparently explains some of the disagreement about the size of the workload effect. Valuable hours were substantially correlated with their SET composite and perceived learning (*r* s about .6; Gillmore & Greenwald, 1994), whereas total hours had only modest correlations with these outcome variables. Hence, it is very likely that their workload variable would have been much more positively correlated with SETs if they had used good hours to infer workload instead of,or in addition to, total hours. Gillmore and Greenwald (1994) also reported that each of their other workload items was positively correlated with their SET composite (*r* s of .14 to .35) and perceived learning (*r* s of .24 to .40). Thus, it appears that these other Workload items would be more internally consistent with good hours than with the total hours and that their a priori model might have better fit their data if good hours were used instead of total hours. In response to this criticism, Greenwald (1998) justified their exclusion of valuable hours, claiming that it could not be modeled appropriately in the models that they posited. We question, however, whether the appropriate decision was to throw out the valuable hours data rather than to throw out the model being tested. The inclusion of valuable hours would, apparently, have substantially undermined support for their interpretations of grade and workload effects. Furthermore, asking students to report both total hours and valuable hours as part of the same survey may subtly change student responses to the total hours item, implicitly encouraging them to place more emphasis on bad hours so as to distinguish between the two variables. In the SEEQ data, for example, the overall teacher and overall course ratings are positively correlated with total hours (*r* s of .18 and .27), and these correlations do not differ much from correlations based on the workload, difficulty, and pace items (*r* s of .22 to .32).

Curiously, Greenwald and Gillmore (1997a) posed an attribution hypothesis and then dismissed it, claiming that it was unable to explain the negative relation between grades and workload. Yet, related to suggestions by Marsh (1980 , 1983 , 1987), if students expect to get poor grades, then the attribution theory predicts that they will attribute their failure to external attributions, such as course difficulty. Hence, in contrast to Greenwald and Gillmore's suggestion, we interpret the negative grade-workload relation as being consistent with their attribution theory hypothesis. Some results of Study 2-the nature of the nonlinear grade effects and the relation between grades and Exams-also seem consistent with attribution theory predictions.

Some prior characteristics, such as prior GPA, year in school,and enrollment have little or no direct effects on SETs but may have moderate indirect effects through grades. Thus, for example, both grades and SETs tend to be higher in more advanced courses than in introductory courses such that part of the grade-SET relation is spurious. If these prior variables are excluded from the analysis, then spurious grade-SET relations that should be attributed to these prior variables are treated as part of the causal effects of grades. Hence, it may only be appropriate to exclude such prior variables if their total effects (not their indirect effects) are negligible and their exclusion does not substantively alter path coefficients associated with other background variables that are retained (e.g., grades and workload). This problem may be evident in the Greenwald and Gillmore (1997b) and Gillmore and Greenwald (1994) analyses. The effect of grades on the composite teacher ratings was modest (b = .18) in the multiple regression that included class size and class level (Gillmore & Greenwald, 1994) but was much larger in the path analyses that excluded these prior background variables. Although there are other differences between the two sets of analyses that may contribute to these different results, the exclusion of prior variables that are related to both grades and SETs will systematically bias paths leading from grades to SETs. For this reason, future analyses of this relation should be much more cautious about excluding these variables when they are available and more cautious about interpretations of results when they are not

available.

More generally, greater emphasis should be placed on the total effects-in addition to the direct effects-of background variables and potential biases. It is inappropriate to exclude background variables because they have no direct effects if they have total effects (i.e., indirect effects that are mediated through intervening variables). Because these mediated effects of prior characteristics are the spurious effects of subsequent variables,the exclusion of prior characteristics leads to a systematic bias in estimated effects of the subsequent variables. This concern is even more critical when there are suppression effects like those for grade and workload. Whereas the positive direct effects of grades and workload on overall teacher ratings were similar in size, the indirect effects of grades were negative such that the total effects of workload were much more positive than those of grades. Direct effects are not dependent upon the ordering of variables in the structural equation model (although they are highly dependent on the inclusion of appropriate prior variables). However, the interpretation of mediated effects-and, thus, total effects-are dependent on whether researchers appropriately identify the ordering of variables in their models. This requires researchers to have a better understanding of their variables so that they can model them appropriately rather than relying on the atheoretical, "dustbowl empiricism" approach of throwing everything into a regression equation. SET research needs to become more theoretical and more rigorous, systematically pursuing a construct validity approach to the interpretation of potential influences on SETs as validity or bias.

Finally, we offer some direction for future research. Studies attempting to show that grade-SET and workload-SET relations (or relations between SETs and any other potential biases) reflect a bias must operationally define bias and grading leniency in a way that is theoretically defensible. More use of the construct validity approach to evaluate interpretations of grade-SET relations is needed (e.g., consideration of prior characteristics,use of multidimensional SETs, teacher self-evaluations, relations with other variables reflecting grading leniency, student achievement, mastery, direct measures of grading leniency). Simple grade-SET correlations are of limited usefulness as there are many conflicting interpretations of these results. Path analytic studies that control for other variables are more useful,particularly when they are designed to test competing explanations of the grade-SET relation. Experimental field studies may be precluded by current ethical standards, but laboratory studies using Dr. Fox-like designs are more promising. More careful consideration of the juxtaposition of grade-SET relations and multisection validity studies (where grading leniency is largely controlled) is needed. Finally, we find it surprising that good quantitative and qualitative blend studies of grade-SET and workload-SET relations have not been pursued.

In summary, our results complement previous research, showing that teachers cannot get higher than average SETs merely by offering easier courses and giving students higher than deserved grades. Indeed, courses demanding the least amount of work tend to receive lower ratings-not higher ratings-and the grade-SET function is relatively flat for grades that are above the mean grade. A detailed evaluation of the implications of the grading leniency hypothesis fails to support it. In contrast to popular myths, the most effective ways for teachers to get high SETs are to provide demanding and challenging materials, to facilitate student efforts to master the materials, and to encourage them to value their learning-in short, to be good teachers. Furthermore,teachers who want to improve their SETs have far more effective and appropriate options available than resorting to counterproductive strategies such as lenient grades and light workloads. In particular, there is ample evidence that SETs and teaching effectiveness can be improved through a cost-effective combination of SET feedback, appropriate consultation, and application of teaching strategies specific to the particular components of teaching effectiveness that teachers choose to target (Marsh & Roche, 1993 , 1994).

Postscript

As a postscript, we pose some questions for researchers who continue to argue in favor of low workload and grading leniency biases. We feel that our research provides at least tentative answers to these questions, thereby undermining much of the argument for such bias interpretations. In response to this article and further research into these controversies, we invite researchers to focus on the following questions and pose new ones so as to better understand these relations:

1. Positive workload-SET correlations: The workload-SET relation is positive except for the highest workload levels. Does this not argue against a low workload bias?

2. Multidimensionality: Grade-SET correlations vary substantially for different SET factors. Does this not argue against a simple bias grading leniency interpretation?

3. Multisection validity studies: Grade-SET *r* s ≈ .5 in multisection validity studies where grading leniency is largely controlled, a relation higher than the typical grade-SET relation. Does this not argue for a validity interpretation?

4. Prior subject interest: Controlling prior subject interest and other prior characteristics substantially reduces grade-SET relations. Does this not argue for a prior characteristics interpretation?

5. Achievement: If controlling for achievement substantially reduces grade-SET correlations, then grade-SET relations do not reflect grading leniency. (Our results suggest this to be the case, but further research with better measures of achievement is needed.) Does this not imply that any claim that grade-SET relations represent grading leniency must control achievement?

6. Interpretation of grades: How can one claim that grades are a satisfactory measure of grading leniency without controlling achievement, prior subject interest, and other prior variables?

7. Nonlinearity grade-SET correlations: The grade-SET function is nearly flat for above-average grades. Does not this argue against a grading leniency interpretation?

8. Grade-SET correlations: Grade-SET relations are small (≈.2 for overall teacher ratings), and much of this small relation is explained by prior characteristics (particularly prior subject interest) and, perhaps, perceived learning. Does this not imply that any remaining effect of grading leniency must be trivially small?

References

Abrami, P. C., d'Apollonia, S. & 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. PsycINFO

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.

Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* (Newbury Park, CA: Sage)

Aleamoni, L. (1987). Student rating myths versus research facts. *Journal of Personnel Evaluation in*

Education, 1, 111-119.
Cashin, W. E. (1988). *Student ratings of teaching: A summary of research* ((IDEA Paper No. 20). Manhattan: Kansas State University, Division ofContinuing Education. (ERIC Document Reproduction Service No. ED 302 567).)
Cashin, W. E. & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology, 84*, 563-572. PsycINFO
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.* (Princeton, NJ: Educational Testing Service)

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* (Hillsdale, NJ: Erlbaum)
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 EducationalResearch Association, Washington, DC. (ERIC Document Reproduction Service No. ED 283 876).)
Covington, M. V. (1997). A motivational analysis of academic life in college.(In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 61-100). New York: Agathon.)
d'Apollonia, S., Lou, Y. & Abrami, P. C. (1998). *Making the grade: A meta-analysis on the influence of grade inflation on student ratings.* (Manuscript submitted for publication)
Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education, 4*, 69-111.
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.)
Feldman, K. A. (1998). Reflections on the study of effective college teaching and student ratings: One continuing quest and two unresolved issues.(In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 13, pp. 35-74). New York: Agathon.)
Franklin, J. & Theall, M. (1996, April). *Disciplinary difference 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' rating data.* (Paper presented at the annual meeting of the American EducationalResearch Association, New York.)
Freedman, R. D. & Stumpf, S. A. (1978). Student evaluation of courses and faculty based on a perceived learning criterion: Scale construction, validation, and comparison of results. *Applied Psychological Measurement, 2*, 189-202. PsycINFO
Frey, P. W. (1978). A two dimensional analysis of student ratings of instruction. *Research in Higher Education, 9*, 69-91. PsycINFO
Gillmore, G. M. & Greenwald, A. G. (1994). *The effects of course demands and grading leniency on student ratings of instruction* ((Report No. 94-4). Seattle: University of Washington, Office ofEducational Assessment.)
Greenwald, A. (1996). *Applying social psychology to reveal a major (but correctable) flaw in student*

evaluations of teaching. (Unpublished manuscript, University of Washington)
Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*, 1182-1186. PsycINFO
Greenwald, A. G. (1998, April). *The positive relationship between course grades and course ratings: What is the cause and what, if anything, should be done about it?* (Symposium conducted at the annual meeting of the AmericanEducational Research Association, New York.)
Greenwald, A. G. & Gillmore, G. M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*, 1209-1217. PsycINFO
Greenwald, A. G. & 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. PsycINFO
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. PsycINFO
Howard, G. S. & Maxwell, S. E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education, 16*, 175-188. PsycINFO
Jöreskog, K. G. & Sörbom, D. (1993). LISREL-8: Structural equation modeling with the SIMPLIS command language.(Chicago: Scientific Software International)
Linn, R. L., Centra, J. A. & Tucker, L. R. (1974). *Between, within, and total factor analyses of students' rating of instruction* ((Report No. RB-74-39). Princeton, NJ: Educational Testing Service)
Marsh, H. W. (1976). *The relationship between background variables and students' evaluations of instructional quality.* (Los Angeles: University of Southern California, Office ofInstitutional Studies.)
Marsh, H. W. (1980). The influence of student, course, and instructor characteristics on evaluations of university teaching. *American Educational Research Journal, 17*, 219-237. PsycINFO
Marsh, H. W. (1982). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement, 6*, 47-59.
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. PsycINFO
Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality,reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707-754. PsycINFO
Marsh, H. W. (1986a). Applicability paradigm: Students' evaluations of teaching effectiveness in different countries. *Journal of Educational Psychology, 78*, 465-473. PsycINFO
Marsh, H. W. (1986b). The self serving effect (bias?) in academic attributions: Its relation to academic achievement and self-concept. *Journal of Educational Psychology, 78*, 190-200. PsycINFO
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.
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. PsycINFO
Marsh, H. W., Balla, J. R. & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes.(In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 315-353). Hillsdale, NJ: Erlbaum.)
Marsh, H. W. & 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. & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course

level. *Teaching and Teacher Education, 7*, 9-18.
Marsh, H. W. & Overall, J. U. (1979). *Validity of students' evaluations of teaching: A comparison with instructor self-evaluations by teaching assistants,undergraduate faculty, and graduate faculty.* ((ERIC Document Reproduction Service No. ED177 205)
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. PsycINFO
Marsh, H. W. & Roche, L. A. (1994). *The use of students' evaluations of university teaching to improve teaching effectiveness.* (Canberra: Australian Government Publishing Service)
Marsh, H. W. & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52*, 1187-1197. PsycINFO
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. PsycINFO
McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe, 65*, 384-397.
McKeachie, W. J. (1997a). Good teaching makes a difference-And we know what it is.(In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 396-411). New York: Agathon.)
McKeachie, W. J. (1997b). Student ratings: The validity of use. *American Psychologist, 52*, 1218-1225. PsycINFO
Michigan State University, Office of Evaluation Services. (1972). *Student Instructional Rating System response and student characteristics* ((SIRS Research Report No. 4). East Lansing: Author)
Perry, R. P. (1997). Perceived control in college students: Implications for instruction in higher education.(In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 11-60). New York: Agathon.)
Pohlman, J. T. (1975). A multivariate analysis of selected class characteristics and student ratings of instruction. *Multivariate Behavioral Research, 10*, 81-91.
Schwab, D. P. (1976). *Manual for the Course Evaluation Instrument.* (Madison: University of Wisconsin, School of Business)
Snyder, C. R. & Clair, M. (1976). Effects of expected and obtained grades on teacher evaluation and attribution of performance. *Journal of Educational Psychology, 68*, 75-82.
Theall, M., Franklin, J. & Ludlow, L. (1990). Attributions and retributions: Student ratings and the perceived causes of performance. *Instructional Evaluation, 11*, 12-17.
Watkins, D. (1994). Student evaluations of teaching effectiveness: A cross-cultural perspective. *Research in Higher Education, 35*, 251-266.
Weiner, B. (1980). *Human motivation.* (New York: Holt, Rinehart & Winston)
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. PsycINFO

This research was funded in part by a grant from the Australian Research Council. We thank Alexander Yeung, Phil Abrami, Bill McKeachie, and Ken Feldman for their helpful comments on earlier versions of this article. We also acknowledge the late Roberta (Bobbie) Vaille for her valuable insights on excellent teaching.
Correspondence may be addressed to Herbert W. Marsh, >Faculty of Education and Languages, University of Western Sydney Macarthur, Campbelltown, New South Wales, Australia, 2560.

Electronic mail may be sent to h.marsh@uws.edu.au
Received: July 2, 1998
Revised: January 5, 1999
Accepted: January 5, 1999

Table 1. Three- and Four-Factor Solutions: Factor Loadings, Path Coefficients, and Factor Correlations

Table 2. Study 2: Polynomial Regression Analyses Relating Background Variables (Linear and Quadratic Components) Separately to Each Student Evaluation Factor (N = 5,433 Classes)

Table 2
Study 2: Polynomial Regression Analyses Relating Background Variables (Linear and Quadratic Components) Separately to Each Student Evaluation Factor (N = 5,433 Classes)

Student evaluation	Background variable					
	GPA	Year	PSI	Enroll	Grade	Workload
Overall teacher						
Linear	.05*	.06*	.23*	−.10*	.20*	.19*
Quadratic	.02	−.03	−.02	.10*	−.07*	−.13*
Overall course						
Linear	.09*	.10*	.38*	−.13*	.25*	.25*
Quadratic	.03	−.04	−.03	.13*	−.06*	−.11*
Learning						
Linear	.17*	.18*	.53*	−.21*	.32*	.17*
Quadratic	.01	−.05*	−.04	.17*	−.04	−.02
Enthusiasm						
Linear	.04	−.01	.15*	.03	.07*	.14*
Quadratic	−.01	.03	−.02	.05*	−.03	−.06*
Organization						
Linear	−.05*	−.04	.06*	−.02	.03	.12*
Quadratic	−.01	−.03	−.05*	.06*	−.09*	−.12*
Group Interaction						
Linear	.15*	.30*	.28*	−.33*	.30*	−.05
Quadratic	.04	−.10*	.05*	.11*	.02	.05*
Individual Rapport						
Linear	.04	.13*	.06*	−.23*	.19*	.03
Quadratic	.07*	−.07*	.04*	.08*	.01	−.09*
Breadth						
Linear	−.01	.01	.07*	.04	.09*	.10*
Quadratic	−.00	−.06*	−.09*	−.00	−.04*	−.10*
Exams						
Linear	.09*	.14*	.17*	−.20*	.31*	.07*
Quadratic	.04	−.05*	.06*	.15*	−.07*	−.09*
Assignments						
Linear	.08*	.05*	.23*	−.07*	.14*	.26*
Quadratic	.03	−.03	−.02	.06*	−.04*	−.17*

Note. Coefficients are standardized beta weights relating each background variable to all SET factors. GPA = prior grade point average; year = mean year in school, 1 (freshman) to 5 (graduate); PSI = prior subject interest; enroll = enrollment; grade = class-average expected grade.
*$p < .001$.

Table 3. Variance Components for Class-Average Expected Grades and Workload Alone (Equation 1) and in Combination With Additional Background Variables (Equations 2, 3, and 4: $N = 5,433$ Classes)

(Equation 1) and in Combination With Additional Background Variables (Equations 2, 3, and 4: N = 5,433 Classes)

Student evaluation	Grade variance component				Work variance component			
	1	2	3	4	1	2	3	4
Overall Teacher								
Grade or workload only	3.93	2.16	3.33	2.47	3.72	2.32	3.19	2.21
PSI		3.70		3.52		4.06		3.40
GPA, year, enroll, enroll quad			1.45	1.28			0.15	0.86
Overall Course								
Grade or workload only	6.25	2.71	4.71	3.06	6.35	3.41	5.36	3.29
PSI		10.90		10.19		11.50		9.79
GPA, year, enroll, enroll quad			2.22	1.52			2.76	1.05
Learning								
Grade or workload only	9.93	3.78	6.03	3.52	3.08	0.61	1.99	0.05
PSI		22.19		19.31		25.87		20.33
GPA, year, enroll, enroll quad			4.66	1.79			7.48	1.94
Enthusiasm								
Grade or workload only	0.54	0.15	0.50	0.23	1.87	1.23	1.66	1.16
PSI		1.83		1.99		1.58		1.74
GPA, year, enroll, enroll quad			0.50	0.65			0.03	0.04
Organization								
Grade or workload only	0.10	0.03	0.30	0.20	1.53	1.31	1.53	1.31
PSI		0.28		0.42		0.13		0.28
GPA, year, enroll, enroll quad			1.30	1.44			1.11	1.26
Group								
Grade or workload only	9.20	5.98	4.50	3.65	0.21	0.98	0.54	0.10
PSI		4.47		2.34		8.46		3.68
GPA, year, enroll, enroll quad			9.27	7.58			14.75	9.97
Individual Rapport								
Grade or workload only	3.62	2.30	2.53	2.55	0.11	0.05	0.04	0.04
PSI		0.02		0.02		0.28		0.00
GPA, year, enroll, enroll quad			4.88	4.88			5.90	5.63
Breadth								
Grade or workload only	0.78	0.55	1.03	0.84	0.96	0.75	1.10	0.87
PSI		0.25		0.49		0.27		0.45
GPA, year, enroll, enroll quad			0.58	0.82			0.46	0.65
Exams								
Grade or workload only	9.83	7.92	7.61	7.00	0.48	0.15	0.19	0.08
PSI		0.95		0.50		2.53		1.00
GPA, year, enroll, enroll quad			4.31	3.86			6.25	4.72
Assignments								
Grade or workload only	1.83	0.68	1.15	0.67	6.66	4.79	5.98	4.66
PSI		4.21		3.79		3.49		2.95
GPA, year, enroll, enroll quad			0.69	0.27			0.69	0.15

Note. Grades and workload were related to each SEEQ score in a series of multiple regressions that included only the target predictor variable (grade or workload) by itself (Equation 1) or the target variable combined with PSI (Equation 2), GPA, year, enrollment, enrollment^2 (Equation 3), or all of these background variables (Equation 4). The variance component for step one is r^2 (e.g., the overall teacher–grade correlation is .198 and the variance component is $.198^2 \times 100\%$), and for each subsequent step it is the change in R^2 that would result in excluding the predictor set from the regression equation. Grade = class-average expected age; PSI = prior subject interest; GPA = prior grade point average; year = mean year in school; enroll = enrollment; quad = quadratic component.

Table 4. Relations of Overall Teacher Ratings With Class-Average Grade Expectation and Workload in Combination With Other Background Variables

Table 4

Relations of Overall Teacher Ratings With Class-Average Grade Expectation and Workload in Combination With Other Background Variables

Predictor variable	B_{fin}	Var	Predictor variable	B_{fin}	B_{init}	Var	R	R^2
Grade	.20*	.04					.20	.04
Grade	.20*	.04	Grade quad	−.04*	−.04	.01	.21	.04
Grade	.20*	.04	GPA	−.02	.05	.00	.20	.04
Grade	.20*	.04	Year	.01	.06	.00	.20	.04
Grade	.15*	.02	PSI	.20*	.23	.04	.28	.08
Grade	.19*	.03	Enroll	−.06*	−.10	.00	.21	.04
Grade	.18*	.03	Enroll	−.07*	−.10	.00	.23	.05
			Enroll quad	.02*	.03	.01		
Work	.19*	.04					.19	.04
Work	.19*	.04	Work quad	−.08*	−.08	.02	.23	.05
Work	.19*	.04	GPA	.02	.05	.00	.19	.04
Work	.19*	.04	Year	.06*	.06	.00	.20	.04
Work	.16*	.02	PSI	.20*	.23	.04	.28	.08
Work	.19*	.04	Enroll	−.09*	−.10	.01	.21	.04
Work	.18*	.03	Enroll	−.09*	−.10	.01	.23	.05
			Enroll quad	.02*	.03	.01		
Work	.28*	.07	Grade	.28*	.20	.07	.33	.11
Work	.28*	.07	Grade	.28*	.20	.07	.34	.11
			Grade quad	−.05*	−.08	.01		

Note. B_{fin} = unstandardized beta weight when all variables are in the final regression equation; B_{init} = unstandardized beta weight when only the one predictor variable is considered; var = variance component (change in R^2 that would result from removing the one predictor variable from the regression equation); grade = class-average expected grade; GPA = prior grade point average; year = year in school; PSI = prior subject interest; enroll = enrollment; quad = quadratic component.

*$p < .001$.

Table 5. Background Effects on Overall Teacher Ratings: Total, Direct, and Indirect Effects and Correlations

[redacted]

Table 6. Overall Teacher Ratings, Workload, and Class-Average Expected Grades: Trends (Means, Standard Deviations, and Correlations) Over a 12-Year Period (N = 5,433)

Table 6

Overall Teacher Ratings, Workload, and Class-Average Expected Grades: Trends (Means, Standard Deviations, and Correlations) Over a 12-Year Period (N = 5,433)

		Teacher		Workload		Grade		Correlations		
Year	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	Teacher with grade	Teacher with workload	Workload with grade
1977	435	−.14	1.05	.03	1.03	.10	0.99	.24	.18	−.24
1978	465	−.07	0.96	.03	0.99	.13	0.94	.15	.20	−.28
1979	483	−.05	1.01	−.05	1.05	.13	0.97	.18	.20	−.30
1980	490	.02	1.02	−.05	0.98	.27	1.01	.24	.14	−.24
1981	499	.01	1.07	−.01	1.03	−.09	1.05	.21	.24	−.35
1982	503	.00	1.01	−.12	1.05	.03	1.00	.19	.21	−.34
1983	503	.03	0.95	−.05	1.01	.00	0.98	.09	.31	−.36
1984	498	−.03	1.06	−.05	1.01	−.04	0.98	.16	.24	−.27
1985	500	.07	0.95	−.01	0.99	−.11	0.99	.21	.18	−.30
1986	509	.04	0.98	.12	0.93	−.19	1.00	.27	.11	−.25
1987	548	.07	0.93	.13	0.90	−.18	0.99	.29	.09	−.25
Total	5433	.00	1.00	.00	1.00	.00	1.00	.20	.19	−.29
		Variance explained by year (%)								
Linear		.25*		1.29*		.10		.01	.03	.00
Quadratic		.02		.02		.35*		.09	.15	.06
Cubic		.03		.03		.01		.02	.05	.00

Note. Results are presented separately for each year and for the total period (1977–1987 academic years). For means and correlations, the linear, quadratic, and cubic effects of year were tested; the percentage of variance explained is presented along with a test of its statistical significance. Teacher = overall teacher ratings; grade = class-average grade expectations.

*$p < .001$.

Figure 1. Structural equation models (Study 1) based on a reanalysis of the correlation matrix published by Greenwald and Gillmore (1997b). Critical differences between the two models are the effects of grade on overall evaluations when perceived learning is hypothesized as one component of the overall evaluation (Figure 1 A) and when it is hypothesized as a measure of student achievement (Figure 1 B) that is a basis for grades,workload, and overall evaluation. Grade = class-average grade expectations; work = workload.

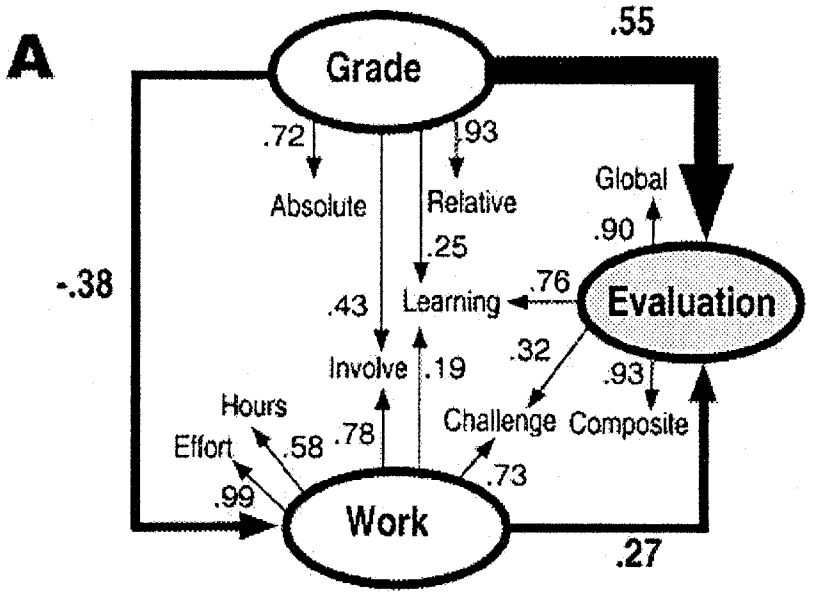


Figure 2. Nonlinear relations between overall teacher ratings and three background variables (see also Table 3): class-average grade expectations (a), workload (b), and enrollment (c). Dashed horizontal and vertical lines are the mean values of each variable.

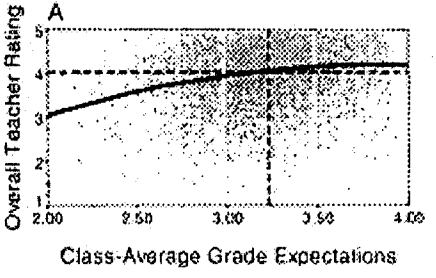
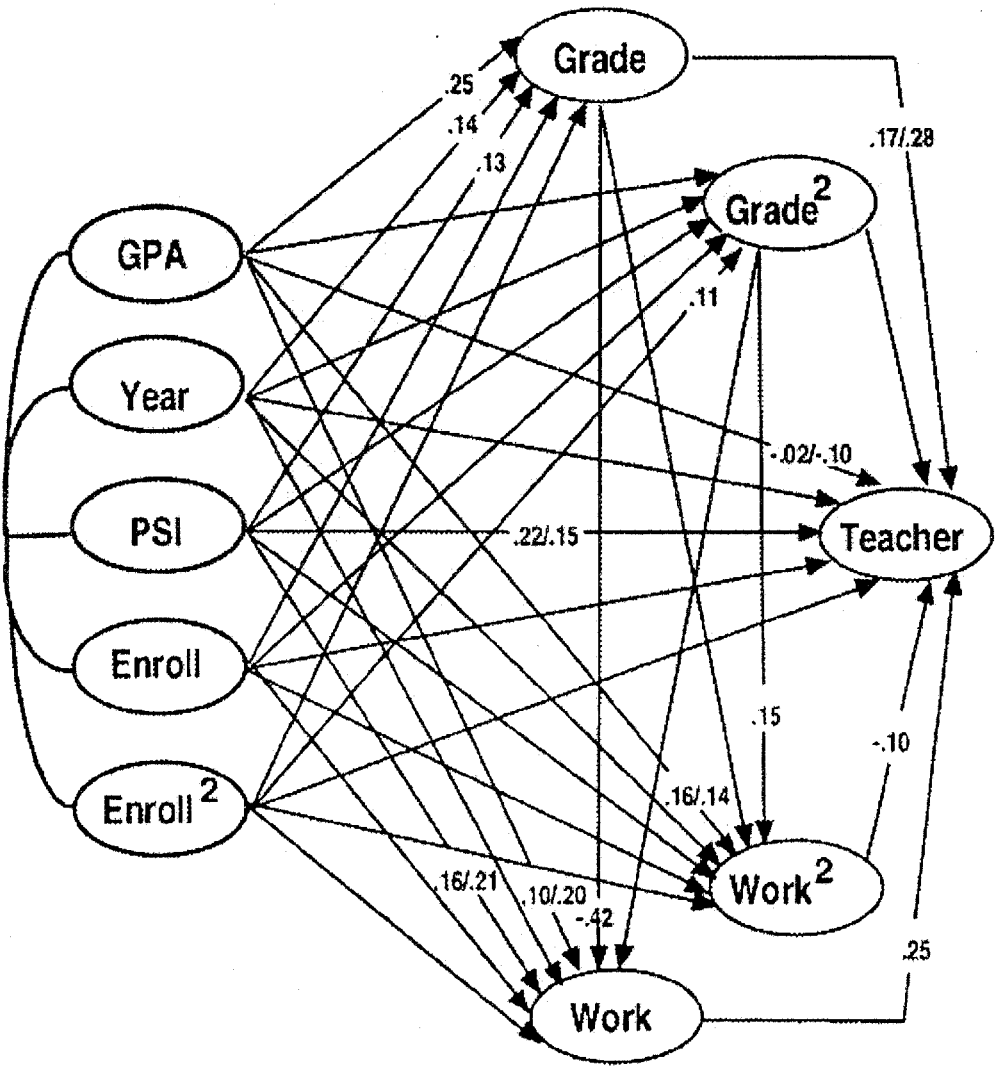


Figure 3. Structural equation model relating class-average grade expectations (grade), workload, and other background variables to overall teacher ratings. Effects are presented whenever total effects (presented first) or direct effects (presented second unless they are the same as total effects) are greater than .10 (all effects are presented in Table 6). GPA = prior grade point average; year = mean year in school; PSI = prior subject interest; grade = classaverage grade expectations; enroll = enrollment; work = workload; teacher = overall teacher rating. Also included are quadratic components of enrollment, grades, and workload.


Grade
Grade^2
GPA
Year
PSI
Enroll
Enroll^2
Teacher
Work^2
Work
.25
.14
.13
.17/.28
.11
-.02/-.10
.22/.15
.15
-.10
.16/.14
.16/.21
.10/.20
-.42
.25