Navigating Student Ratings of Instruction

Article in American Psychologist · November 1997
DOI: 10.1037/0003-066X.52.11.1198

CITATIONS
229

READS
337

2 authors, including:

Philip C. Abrami
Concordia University Montreal
91 PUBLICATIONS 3,002 CITATIONS

All in-text references underlined in blue are linked to publications on ResearchGate, letting you access and read them immediately.

Available from: Philip C. Abrami
Retrieved on: 11 October 2016
Many colleges and universities have adopted the use of student ratings of instruction as one (often the most influential) measure of instructional effectiveness. In this article, the authors present evidence that although effective instruction may be multidimensional, student ratings of instruction measure general instructional skill, which is a composite of three subskills: delivering instruction, facilitating interactions, and evaluating student learning. The authors subsequently report the results of a meta-analysis of the multisection validity studies that indicate that student ratings are moderately valid; however, administrative, instructor, and course characteristics influence student ratings of instruction.

In navigating oceans, sailors depend on a variety of nautical instruments. Their success in traversing poorly charted seas is founded on the knowledge that the different instruments measuring latitude will provide similar readings of latitude but different readings of longitude. Similarly, when practitioners interpret student ratings, they expect different measures of the trait in question (i.e., instructional effectiveness) to give substantially similar results. However, they expect measures of irrelevant traits (e.g., athletic prowess, grading leniency, charisma) to give substantially different results. That is, they expect student ratings of instruction to reflect the true impact of teaching on student learning.

Student ratings of instruction, first introduced into North American universities in the mid-1920s (Doyle, 1983), have been the subject of a huge body of literature including primary studies, reviews, and books. This literature has dealt not only with the psychometric properties of student ratings (Costin, Greenough, & Mengers, 1971; Doyle, 1983; Feldman, 1978; Marsh, 1984) and the factor structure of ratings (Kulik & McKeachie, 1975; Marsh, 1991a, 1991b; McKeachie, 1979) but also with practical guides to faculty evaluation (Arreola, 1995; Braskamp & Ory, 1994; Centra, 1993). Many researchers have concluded that the reliability (Centra, 1993) and the validity (Abrami, d’Apollonia, & Cohen, 1990; Cohen, 1981; Feldman, 1989, 1990; Marsh, 1987) of student ratings are generally good and that student ratings are the best, and often the only, method of providing objective evidence for summative evaluations of instruction (Scriven, 1988). However, a number of controversies and unanswered questions persist. Our objective in this article is to briefly and selectively review some of the literature on student ratings of instruction that addresses the following three questions: What is the structure of student ratings of instructional effectiveness? Is instructional effectiveness substantially correlated with other indicators of instructor-mediated learning in students? To what extent do student ratings confound irrelevant variables with instructional effectiveness?

What Is the Structure of Student Ratings of Instructional Effectiveness?

Most postsecondary institutions have adopted student ratings of instruction as one (often the most influential) measure of instructional effectiveness. Student rating forms are paper-and-pencil instruments on which students indicate their responses to items on some numerically based scale. Specific items such as “Does the instructor speak clearly?” and “Rate the degree to which the instructor was friendly toward students” tap specific instructional dimensions such as presentation skills and rapport. Global items such as “How would you rate the overall effectiveness of the instructor?” tap students’ overall impressions of the instructor. Omnibus student rating forms, such as the Student Instructional Rating System (Michigan State University, 1971), the Students’ Evaluation of Educational Quality (SEEQ; Marsh, 1987), and the Student Instructional Report (Linn, Centra, & Tucker, 1975), are usually multidimensional, consisting of several items tapping specific dimensions. They also may include overall global items assessing students’ overall impressions of the course, the instructor, and the value of the course.

Characteristics of Effective Instruction

Items on student rating forms reflect the characteristics that experts believe (a) can be judged accurately by students and (b) are important to teaching. Nevertheless, researchers define instructional effectiveness from a num-

Sylvia d’Apollonia and Philip C. Abrami, Centre for the Study of Learning and Performance, Concordia University, Montreal, Quebec, Canada.

The research reported in this article was supported by grants from the Social Sciences and Humanities Research Council (Government of Canada) and Fonds pour la formation de chercheurs et l’aide à la recherche (Government of Quebec).

Correspondence concerning this article should be addressed to Sylvia d’Apollonia, Centre for the Study of Learning and Performance, Concordia University, 1455 West DeMaisonneuve Boulevard, Montreal, Quebec, Canada H3G 1M8. Electronic mail may be sent via Internet to sdapoll@vax2.concordia.ca.
ber of different perspectives (Abrami, d’Apollonia, & Rosenfield, 1996). These definitions may focus on either aspects of the instructional process (e.g., preparation of course materials, provision of feedback, grading) or the products that effective instruction promotes in students (e.g., subject-matter expertise, skill in problem solving, positive attitudes toward learning). However, in the past 30 years, faculty members have begun to use a variety of teaching methods other than didactic teaching. These alternative methods range from interactive seminars, laboratory sessions, and cooperative learning to more independent methods such as computer-assisted instruction, individualized instruction, and internships. Whether the definitions of instructional effectiveness are based on instructional products, processes, or the causal relationships between the two, they do not necessarily generalize across these other instructional contexts.

Many multidimensional student rating forms contain items that assess the instructor’s beliefs, knowledge, attitudes, organization and preparation, ability to communicate clearly and to stimulate students, interaction with students, feedback, workload, and fairness in grading. Feldman (1976, 1983, 1984) reviewed studies that either investigated students’ statements about superior instructors or correlated students’ evaluations of instructors with instructors’ characteristics. Feldman developed several schemas consisting of high-inference categories that subsequently have been used to code items in different student rating forms (Abrami, 1988; Abrami & d’Apollonia, 1990; Marsh, 1991a, 1991b).

Abrami and d’Apollonia (1990) explored the consistency and uniformity of the student rating forms used in multisection validity studies by coding and analyzing the items using the aforementioned coding schema. They concluded that student rating forms, each purported to measure instructional effectiveness, were not consistent in their operational definitions of instructional effectiveness. Thus, no one rating form represents effective instruction across contexts.

**Factor Studies of Individual Student Rating Forms**

A number of student rating forms have been factor analyzed to determine their dimensional structure (Frey, Leonard, & Beatty, 1975; Linn et al., 1975; Marsh, 1987). For example, Marsh and Hocevar (1984, 1991) identified nine instructional dimensions on the SEEQ and demonstrated that the factor structure of the SEEQ was invariant across different groups of students, academic disciplines, instructor levels, and course levels. However, the generalization of the factor structure across different pedagogical methods (e.g., lectures, small discussion groups) was not explored.

Some researchers, notably Marsh and his colleague (Marsh, 1983a; Marsh & Hocevar, 1984, 1991), have interpreted the factor invariance of student rating forms as providing evidence for the construct validity of distinct instructional dimensions. However, Abrami and d’Apollonia (1990, 1991), Abrami et al. (1996), and Feldman (1976) have criticized the interpretation of factor analysis as evidence of the construct validity of student ratings. They contended that there were a number of methodological problems with this use of factor analysis.

First, factor solutions reflect the items in the specific rating instrument in question; they do not necessarily reflect the construct in question (i.e., instructional effectiveness). Second, analysts often extract all factors having eigenvalues greater than one. Zwick and Velicer (1986) reported that this practice seriously overestimates the number of factors and extracts minor factors at the expense of the first principal component. Finally, the choice of factor analytical method determines the factor solution. For example, factor analysis conducted without rotation is designed to resolve one general or global component, whereas factor analysis conducted with rotation is designed to resolve specific factors of equal importance (Harman, 1976). However, the factor solutions are indeterminate (Gorsuch, 1983); many solutions exist, each resolving exactly the same amount of the total variance. Therefore, the solution that best describes the construct cannot be determined solely empirically. The solution must be supported by theory and by empirical research other than factor analysis.

Abrami and d’Apollonia (1991) reanalyzed the factor structure of the SEEQ by reconstructing the reproduced correlation matrix from the published pattern and factor correlation matrices (Marsh & Hocevar, 1984, 1991). They subsequently factor analyzed the estimated correlation matrix and replicated Marsh and Hocevar’s (1984, 1991) results within rounding error. They argued that because 31 of the 35 items loaded heavily (> .60) on the first principal component, the principal-components solution was interpretable and provided evidence of a Global Instructional Skill factor. The two global items (Overall Instructor Ratings and Overall Course Ratings) had the highest loadings (.94) on the first principal component and explained almost 60% of the total variance in ratings. The remaining four items, concerning course difficulty and workload, loaded heavily on the second principal component. This component explained an additional 11% of the variance.

The results of similar reanalyses of the factor studies for seven other multidimensional student rating forms are presented in Table 1. The principal-components solutions indicate that for six of the seven student rating forms, there was a large global component with 30% or more of the variance in student responses resolved by this first principal component. In general, most of the factor studies can be interpreted as providing evidence for one global component rather than several specific factors.

Thus, the factor studies do not provide incontestable evidence that student rating forms measure distinct instructional factors. An alternative, and mathematically equivalent, interpretation is that student rating forms measure a global component, General Instructional Skill. Factor analysts may rotate the principal-components solution and reify specific factors, but these specific factors may
Table 1
Principal-Components Solution of Student Rating Forms

<table>
<thead>
<tr>
<th>Factor study</th>
<th>% of variance extracted by principal components, in order of extraction (k &gt; 1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linn, Centra, &amp; Tucker (1975)</td>
<td>14.3 11.1 9.2 8.6 5.9 4.3</td>
</tr>
<tr>
<td>Bolton (personal communication, October 7, 1990); Bolton, Bonge, &amp; Marr (1979)</td>
<td>45.1 10.1 8.2 7.6 5.9</td>
</tr>
<tr>
<td>Michigan State University (1971)</td>
<td>34.5 11.7 10.5 8.5 6.5</td>
</tr>
<tr>
<td>Frey, Leonard, &amp; Beatty (1975)</td>
<td>39.9 14.5 9.5 7.2 6.5 5.1</td>
</tr>
<tr>
<td>Hartley &amp; Hogan (1972)</td>
<td>39.2 5.7 4.4</td>
</tr>
<tr>
<td>Mintzes (1977)</td>
<td>30.0 7.5 6.7 5.9 5.2</td>
</tr>
<tr>
<td>Pambookian (1972)</td>
<td>33.4 26.1 9.7 7.8 6.6 5.2</td>
</tr>
</tbody>
</table>

be artifacts of the analytical procedures. The selection of which interpretation (specific factors or general skill) reflects instructional effectiveness as perceived by students cannot be made on the basis of further factor analyses, confirmatory or otherwise. Rather, this choice must be based on theory and evidence using other independent analytical methods.

Information-Processing Models of Student Ratings

Information-processing models of performance ratings (Cooper, 1981; Fiske & Neuberg, 1990; Judd, Drake, Downing, & Krosnick, 1991; Murphy, Jako, & Anhalt, 1993; Nathan & Lord, 1983) suggest that individuals minimize cognitive load by using common prototypes to organize their schemas of persons, roles, and events. Raters use supraordinate features, like general impressions, to attend to, store, retrieve, and integrate judgments of specific behaviors. To the degree that items on rating forms are semantically similar, they function as retrieval cues, activating the same node in the rater's associative network. Furthermore, because persons, roles, and behaviors are stored together, once an individual's personality schema is activated, roles can infer specific behaviors, or vice versa (Trzebinski, 1985). B. E. Becker and Cardy (1986) and Cooper (1981) argued that this reliance on general impressions (halo effects) not only does not distort ratings but also may increase accuracy. Thus, according to information-processing models of performance ratings, factor studies of student ratings reflect not only the structure of instructional effectiveness but also students' cognitive processes while they are rating their instructors.

Researchers also have shown that student ratings collected in field settings demonstrate that students rate specific dimensions of instruction on the basis of their global evaluation. Marsh (1982, 1983b) reported that the halo effect accounted for 19% of the variance in student ratings. Whitely and Doyle (1976) demonstrated in a classroom setting that the factor structure of student ratings of instruction corresponded to students' valid implicit theories of teaching. Similarly, Murray's (1984) study indicated that students' global ratings of instructors were accurate when compared with trained observers' ratings of specific behaviors of instructors.

Abrami and d'Apollonia (1990, 1991) suggested that the general question that needs to be addressed is the dimensionality of instructional effectiveness across different forms and different types of classes. That is, the important question isn't whether one student rating form has a robust factor structure but rather whether the construct of instructional effectiveness has a common meaning or structure across different forms and contexts. This question has been approached both logically and empirically.

Logical approaches to integration of factor studies. Several researchers (Feldman, 1976; Kulik & McKeachie, 1975; Marsh, 1987, 1991a, 1991b; Widlak, McDaniel, & Feldhusen, 1973) logically integrated factor studies across forms purporting to measure the same dimensions of effective instruction. Feldman as well as Widlak et al. concluded that across multidimensional student rating forms, teaching is evaluated in terms of three roles: presenter, facilitator, and manager.1

Empirical approaches to integration of factor studies. Abrami and his colleagues (Abrami et al., 1996; d'Apollonia, Abrami, & Rosenfield, 1993) used multivariate approaches to meta-analysis to uncover the common factor structure across student rating forms. They coded the 458 items in 17 student rating forms into common instructional categories, extracted a reliable set of intercategory correlation coefficients from the reproduced correlation matrices, aggregated them to produce an aggregate correlation matrix, and subsequently factor analyzed that correlation matrix. They reported that there

---

1 Chau (1994) conducted a confirmatory higher order factor analysis on responses to the SEEQ and also demonstrated the presence of the same three factors.

2 Intercategory correlation coefficients for 195 items were eliminated from the data set because the values were not consistent (homogeneous) across student rating forms.
was a large common principal component across student rating forms that explained about 63% of the variance in instructional effectiveness. When the factor solution was rotated obliquely, four first-order factors were obtained. The first factor, in order of importance, was tentatively identified as representing the instructor's role in delivering information. It included the global scales (Overall Course and Instructor Ratings) as well as such behaviors as monitoring of learning, enthusiasm for teaching, clarity, presentation, and organization. The second factor was tentatively identified as representing the instructor's role in facilitating a social learning environment. It included such behaviors as general attitudes, concern for students, availability, respect for others, and tolerance of diversity. The third factor was tentatively identified as representing the instructor's role in evaluating student learning. It included such behaviors as evaluation, feedback, and the provision of a friendly classroom atmosphere. The fourth factor included four miscellaneous teaching behaviors: maintaining discipline, choosing appropriate required materials, knowing the content, and delivering information. It included the global scales and the principal-components analysis indicated that student ratings of instruction measure General Instructional Skill. General Instructional Skill consists of three subskills: delivering instruction, facilitating interactions, and evaluating student learning.

In conclusion, Marsh and his colleague (Marsh & Hocevar, 1984, 1991) contended that the factor invariance of well-designed rating forms, such as the SEEQ, supports the construct validity of specific instructional dimensions. However, an equally viable interpretation of the factor studies is that student rating forms measure a global component, General Instructional Skill. This interpretation is supported by three independent lines of evidence: secondary factor analyses, studies on human information processing during performance ratings, and the integration of factor studies.

**Is Instructional Effectiveness Correlated With Other Indicators of Student Learning?**

One important criterion for the validity of student ratings of instruction is that the teaching processes reflected in student rating forms should cause (or at least predict) student learning. That is, if student ratings measure instructional effectiveness, students whose instructors are judged the most effective also should learn more. One of the methods for determining this aspect of the validity of student ratings is the multisection validity design.

**Multisection Validity Design**

The multisection validity design, first used by Remmers, Martin, and Elliot (1949), is used by researchers to correlate the section average score on student ratings with the section average score on a common achievement test across multiple sections of a course. In the strongest multisection designs, all sections use a common textbook and a common syllabus. Ideally, either students are randomly assigned to sections or student ability is statistically controlled. Although this validation design has been criticized by Marsh (1984, 1987; Marsh & Dunkin, 1992) on a number of grounds (problems of statistical validity, problems with the measurement of both instructional effectiveness and student learning, and problems with rival explanations), this design minimizes the extent to which the correlation between student ratings and achievement can be explained by factors other than instructor influences. Abrami et al. (1996) argued that although the multisection validity design is not perfect, it has high internal validity. To date, more than 40 studies have utilized the multisection validity design to determine the validity of student ratings. In addition, there have been numerous reviews, both qualitative and quantitative. Because this validation design has been used so extensively, with many student rating forms under diverse conditions, it provides the most generalizable evidence for the validity of student ratings.

**Meta-Analysis of Multisection Validity Studies**

The multisection validity literature has been extensively reviewed both qualitatively (Feldman, 1978; Kulik & McKeachie, 1975; Marsh, 1987) and quantitatively (Abrami, 1984; Abrami et al., 1990; Cohen, 1981, 1982; d'Apollonia & Abrami, 1996; Dowell & Neal, 1982; Feldman, 1989; McCallum, 1984). However, not only do the primary researchers reach opposite conclusions about the validity of student ratings, but so do some of the reviewers. For example, Cohen (1981) concluded that student ratings were valid measures of teaching effectiveness, whereas Dowell and Neal and McCallum concluded that ratings, at best, were poor measures of teaching effectiveness.

Abrami, Cohen, and d'Apollonia (1988) explored the reasons for the different conclusions. They compared six published meta-analyses of the multisection validity studies (Abrami, 1984; Cohen, 1981, 1982, 1983; Dowell & Neal, 1982; McCallum, 1984) and identified a number of decisions made by the reviewers that biased their findings. For example, reviewers differed on their inclusion criteria, on the number of outcomes they extracted, and on the analytical techniques they used. They recommended that future quantitative reviewers should improve both the coding of study features and the analysis.
Subsequently, d’Apollonia and Abrami (1996) conducted a meta-analysis in which they extracted all the validity coefficients from 43 multisection validity studies (N = 741), coded them on the basis of the aforementioned common factor structure across rating forms and used multivariate methods of meta-analysis (Raudenbush, Becker, & Kalaian, 1988) to model the dependencies among the data and to calculate population parameters. They reported that the mean validity coefficient of student ratings of general instructional skill across the 43 studies was .33, with the 95% confidence interval extending from .29 to .37.3

However, the correlation between student ratings and student learning was attenuated by unreliability in both the rating and achievement instruments. The average reliabilities of the student rating and achievement instruments in the multisection validity studies were .74 and .69, respectively. Therefore, when the correlation coefficient between student ratings of general instructional skill and student learning was corrected for attenuation (Downie & Heath, 1974), the correlation became .47, with a 95% confidence interval extending from .43 to .51. Thus, there was a moderate to large association between student ratings and student learning, indicating that student ratings of General Instructional Skill are valid measures of instructor-mediated learning in students. The variability in the validity coefficients reported by the primary investigators can be explained, in large part, by differences in the sampling variance of the individual studies.

Do Student Ratings Confound Irrelevant Variables With Instructional Effectiveness?

Some critics of student ratings argue that student ratings are not valid indices of instructor-mediated learning in students because irrelevant (biasing) characteristics are correlated with ratings. Validation techniques based on construct-validity approaches, such as multitrait-multimethod approaches, assume that characteristics that reduce the discriminant validity of an instrument are irrelevant. However, the characteristics that covary with student ratings may be relevant, acting as intermediate variables. For example, if student ratings were correlated with grades, some critics (e.g., Greenwald & Gillmore, in press) contend that student ratings would be biased because students might be rewarding their instructors for lenient grading practices. However, as discussed by Abrami et al. (1996), Abrami and Mizener (1985), Marsh (1980, 1987), and Feldman (in press), this interpretation is correct only if student learning is not causally affected by the putative biasing variable. If instructors were grading students leniently, but this practice had no effect on learning, grading leniency would be a biasing variable. Alternatively, if grading leniency enhanced students’ perceptions of efficacy, encouraged students to work harder, or facilitated student learning, grading leniency would not be irrelevant or a biasing variable. Although many variables have been shown to influence student ratings of instruction, unless they can be shown to moderate the validity coefficient (the correlation between student ratings and student learning), they cannot be described as biasing variables.

Cohen (1981, 1982) reported that three study features (instructor experience, timing, and evaluation bias) moderated the validity coefficient of student ratings and, therefore, may bias student ratings. However, Abrami et al. (1988) suggested that the failure to identify more biasing variables may have been the result of both the incomplete identification of possible moderators and too low statistical power. Therefore, they recommended that additional meta-analyses be conducted that address these two issues. Subsequently, d’Apollonia and Abrami (1996, 1997) used nomological coding (Abrami et al., 1990) and more sophisticated meta-analytical techniques to conduct a more thorough meta-analysis of the multisection validity studies. They extracted all the validity coefficients from 43 multisection validity studies, coded them on the basis of the degree to which they represented general instructional skill, and coded 39 study features.

d’Apollonia and Abrami (1996, 1997) reported that 23 individual study features representing methodological and implementation characteristics, instrument characteristics (rating forms and achievement tests), and explanatory characteristics (instructors, students, and settings) significantly moderated the validity of student ratings of instruction. For example, the research quality of the study examining the strength of the association between instructors’ effectiveness and student learning influenced the size of the correlation. Studies in which instructors graded their own students’ examinations, in which in-
in which prior differences in student ability were not controlled reported lower validity coefficients than studies that controlled for these possible sources of bias. In addition to these study-quality variables, the rank, experience, and autonomy of the instructors and the discipline and size of the multisection course also biased the validity of student ratings of instruction.

d'Apollonia and Abrami (1996, 1997) subsequently constructed a hierarchical multiple regression model consisting of the following study features, entered in this sequence: structure (Factor 1: Presenting Material, Factor 2: Facilitating Interaction, Factor 3: Evaluating Learning), source of study (theses or articles), timing (after or before), group equivalence (none, statistical control, or experimental control), and rank (teaching assistants or faculty members). Because many of the primary studies did not report the values for some of these study features (e.g., only about 50% of the primary studies reported the timing of the evaluation), a subset of the data (225 outcomes) for which complete information was given on these features was selected. The goodness-of-fit statistic \( \chi^2 \) was 210.5 \((df = 200)\), indicating that the aforementioned study features explained all the systematic variation in the subset of data.

In conclusion, the published multisection validity literature suggests that under appropriate conditions (all instructors are faculty members, evaluation is carried out prior to students' knowing their final grade, sections are equivalent in terms of student prior ability or equivalence is experimentally controlled) and the validity coefficient is corrected for attenuation, more than 45% of the variation in student learning among sections can be explained by student perceptions of instructor effectiveness. This 45% figure is, of course, an estimate of validity under the above circumstances. In practice, with the usual reliability of measures and the ratings context, only partially matching the set of conditions, the validity coefficient will be different. In any case, the data set is homogeneous, indicating that across different students, courses, and settings, student ratings are consistently valid.

**Policy Implications and Recommendations**

Many experts in faculty evaluation consider that the validity of summative evaluations based on student ratings is threatened by inappropriate data collection, analysis, reporting, and interpretation (Arreola, 1993; Franklin & Theall, 1990; Theall, 1994). The research described in this article indicates that student ratings of instruction measure General Instructional Skill; that interpretations of general instructional skill are substantially correlated with student learning; and that with the exception of a few variables that influence the quality of evaluations (structure, timing, section equivalence, etc.), student ratings of instruction are not plagued with biasing variables.

In this section, we briefly discuss policy implications on the use of student ratings of instruction for summative evaluation.

**Use of Student Rating Forms for Summative Faculty Evaluation**

Although most researchers agree that teaching is multidimensional, they disagree on whether global or specific ratings should be used for summative decisions on retention, tenure, promotion, and salary. Some researchers (Frey et al., 1975; Marsh, 1984, 1987) have recommended that summative evaluations be based on a set of specific scores derived from multidimensional rating forms. They argue that because teaching is multifaceted, it can be evaluated best by measures that reflect this dimensionality. Because there is little agreement on which pedagogical behaviors necessarily define effective instruction, the specific factors in the set can be differentially weighted to meet individual needs.

Other researchers (Abrami, 1985, 1988, 1989; Abrami & d'Apollonia, 1990, 1991; Cashin & Downey, 1992; Cashin, Downey, & Sixbury, 1994) have argued, equally vigorously, that global ratings (or a single score representing a weighted average of the specific ratings\(^4\)) are more appropriate. They have argued that different student rating forms assess different dimensions of effective instruction and that because specific ratings are less reliable, valid, and generalizable than global ratings, they should not be used for personnel decisions.

The factor analysis across student rating forms indicates that specific dimensions are not distinct. Rather, they are chunked during rating into three factors representing three instructional roles. Moreover, these factors are highly correlated and can be represented by a single hierarchical factor—General Instructional Skill. Thus, the research summarized in this article supports the view that global ratings or a single score representing General Instructional Skill should be used for summative evaluations. Such global assessments of instructional effectiveness are judgmental composites "of what the rater believes is all of the relevant information necessary for making accurate ratings" (Nathan & Tippins, 1990, p. 291). Moreover, specific ratings of instructors' effectiveness add little to the explained variance beyond that provided by global ratings (Cashin & Downey, 1992; Hattie & Raviv, 1993). Therefore, we recommend that summative evaluations be based on a number of items assessing overall effectiveness of instructors.

**Best Administrative Practices**

When student ratings are to be used for summative decisions, it is important that the circumstances under which student ratings are collected and analyzed be appropriate

---

\(^4\) Marsh (1991a, 1991b, 1994, 1995) agreed with Abrami that a weighted average of specific dimensions is a good compromise. However, there is no consensus on how to weight the specific dimensions.
and rigorously applied. Although the composition of student rating forms explains a large amount (approximately 45%) of the systematic variation in the validity of student ratings, items must be chosen with care. Because specific items have lower validity coefficients and may not generalize across different instructional contexts, we recommend that short rating forms with a few global items be used. Such scales have higher validity coefficients than most scales containing a large number of heterogeneous items.

The association between student ratings and student learning was also significantly higher when instructional evaluation was carried out after, rather than before, the final examination. Students may be rewarding instructors who have given them high grades, or they may be using grading or their performance on the final examination as one of the indicators of instructional effectiveness. In either case, the relationship between instructional behavior and student learning may be confounded. Thus, because of this ambiguity, we recommend that student ratings be collected before the final examination. However, other administrative conditions, such as student anonymity, the stated purpose of the evaluations, or whether the instructors themselves carried out the evaluation procedure, did not significantly influence the validity of student ratings of instruction. Nevertheless, for ethical and legal reasons, standardized procedures should be used.

Controlling for Bias in Student Ratings of Instruction

The consensus has been that biasing variables play a minor role in student ratings of instruction (Marsh, 1987; Murray, 1984). However, the research summarized in this article indicates that instructors' rank, experience, and autonomy; the course discipline and class size; and section inequivalence moderate the validity of student ratings. In addition, some researchers have suggested that student motivation, course level, instructors' grading leniency, and instructors' expressivity may also bias student ratings of instructors. For example, students rate instructors more accurately when they evaluate full-time faculty teaching large science courses as compared with when they evaluate teaching assistants teaching medium-sized classes in language courses. Similarly, the validity of student ratings is higher when students are randomly assigned to sections than when there are preexisting section inequivalences in student ability. Because instructors have no control over variables such as class size and student registration, some researchers suggest that rating scores be either norm referenced or statistically controlled for putative biasing variables.

Cashin (1992) and Aleamoni (1996) argued that student ratings of instruction should be interpreted relative to those of a reference group teaching similar students in similar courses in the same department. This solution has been criticized by Hativa (1993), Abrami (1993), and McKeachie (1996) on a number of grounds. First, how does one decide on which variables to base the reference groups? Should only the student ratings of instructors who are equally lenient graders (and presumably equally poor instructors) be compared? Second, although several variables have been shown to influence ratings in correlational studies (e.g., grading leniency and instructors' expressivity), these variables may exert their influence through student learning. Therefore, they would not be biasing variables. Third, in many cases, the norm group would be so small that the data would be unreliable. Moreover, if norms are used to rank instructors, these individual differences in rank may be too small to have any practical significance. Finally, Abrami and McKeachie argued that such norms have negative effects on faculty members' morale. By definition, half the faculty would be below the norm, yet they could be excellent teachers.

Other researchers (Greenwald & Gillmore, in press) have suggested that student ratings should be statistically controlled for grading leniency. However, some of the same arguments as those made against norm referencing also hold for statistical control. In effect, statistically controlling for a variable such as grading leniency computes the rating score as if all instructors graded their students with equal lenience. It achieves mathematically what norming does by selection.

Although the aforementioned variables appear to bias student ratings, the multisection validity studies, like other correlational studies, do not indicate whether the influence of the biasing variable is on student ratings or on student learning. For example, science students' global ratings of instruction have a relatively higher validity than those of art students because of problems either with student ratings or with achievement tests. Art students rate their instructors relatively more favorably (Centra & Creech, 1976; Feldman, 1978; Neumann & Neumann, 1985) than do science students. It may be that art students give their instructors undeserved high ratings. Alternatively, achievement tests in art courses may have relatively greater problems with validity and reliability and may not accurately measure what art students have learned. In the first case, student ratings would be biased. However, in the second case, the measurement of student learning would be biased, and student ratings may accurately reflect instructional effectiveness. Therefore, researchers must investigate the mechanisms whereby these putative biasing variables influence student ratings before attempting to control for their influence.

Laboratory studies that have used the "Dr. Fox" paradigm have shown that instructors' expressivity (Abrami, Leventhal, & Perry, 1982; Abrami, Perry, & Leventhal, 1982) and grading practices (Abrami, Perry, & Leventhal, 1982) can unduly influence student ratings of instruction. However, these studies also have indicated that these variables do not seriously bias student ratings. Expressivity has a practically meaningful influence on student ratings of instructors, with high-expressive instructors scoring about 1.20 standard deviations above low-expressive instructors. However, instructors' expressivity
also influences student learning. Researchers who have investigated instructors’ expressivity have concluded that it is not a biasing variable but rather that it exerts its influence by affecting student learning (Murray, Ruston, & Paunonen, 1990). Liberal grading practices increased student ratings, at most, by slightly less than 0.5 on a 5-point scale. Moreover, in some cases they decreased ratings. Thus, grading practices are not a practical threat to the validity of student ratings. In both cases, controlling for the putative biasing variable “corrects” the influence of the putative biasing variable, removing some of the genuine influence of effective instruction from the rating scores. In effect, poor instructors may be doubly rewarded both by students and by evaluators. Instead, we recommend that student ratings not be overinterpreted. In general, experts recommend that comprehensive systems of faculty evaluation be developed, of which student ratings of instruction are only one, albeit important, component (Arreola, 1995; Braskamp & Ory, 1994; Centra, 1993). Within such a system, student ratings should be used to make only crude judgments of instructional effectiveness (exceptional, adequate, and unacceptable).

REFERENCES


Mintzes, J. J. (1977). Field test and validation of a teaching evaluation
We agree with Marsh and Roche (1997, this issue) that teaching but not under the high-incentive conditions typical of college ratings is not consistent, we consider it to be inappropriate to classes. Because the influence of grading practices on student practices influenced student ratings under low-incentive conditions dent ratings of high-expressive instructors but decreased student variables. For example, lenient grading practices increased stu- though grading standards influenced student ratings, both the variables on the validity coefficient of student ratings address our own meta-analyses in this article (see also d’Apollonia, Neumann, & Neumann, Y. (1985). Determinants of students’ instruc- tional evaluation: A comparison of four levels of academic areas. Journal of Educational Psychology, 78, 152–158. Pambookian, H. S. (1972). The effect of feedback from students to college instructors on their teaching behavior. Unpublished doctoral dissertation, University of Michigan.

Comment on Greenwald
Greenwald (1997b, this issue) states that discriminant validity measures the degree to which ratings are not “influenced by factors other than teaching effectiveness” (p. 1182). Thus, such variables must be independent of teaching effectiveness. We consider that studies that investigate the influence of biasing variables on the validity coefficient of student ratings address the question of discriminant validity. Thus, we disagree with Greenwald (1997b) and maintain that Cohen’s (1981, 1982) and our own meta-analyses in this article (see also d’Apollonia, Abrami, & Rosenfield, 1993) that measured the degree to which putative biasing variables distort the validity coefficient of student ratings are as concerned with discriminant validity as they are with convergent validity.

We disagree on two counts with the conclusion made by Greenwald et al. (1997) that there is a moderate to large influence of grading practices on student ratings. First, Abrami, Dickens, Perry, and Leventhal (1980) and Marsh (1987) criticized the field studies included in Greenwald’s (1997a) meta-analysis for not controlling for viable rival hypotheses. Second, one of the shortcomings of the meta-analytical approach used by Greenwald (1997a) is that moderate to large average effect sizes are misleading because the approach overlooks real differences in effect sizes (B. J. Becker, 1992). Abrami et al. (1980) demonstrated in several well-controlled laboratory studies that although grading standards influenced student ratings, both the degree and the direction of the influence depended on contextual variables. For example, lenient grading practices increased student ratings of high-expressive instructors but decreased student ratings of low-expressive instructors. Similarly, grading prac- tices influenced student ratings under low-incentive conditions but not under the high-incentive conditions typical of college classes. Because the influence of grading practices on student ratings is not consistent, we consider it to be inappropriate to report an overall effect size for grading leniency.

Comment on Marsh and Roche
We agree with Marsh and Roche (1997, this issue) that teaching is multidimensional, as are well-designed student rating forms such as the SEEQ. However, students are neither trained raters nor psychometricians. When students are rating teaching, especially retroactively, they tend to rate specific dimensions of instruc- tion on the basis of their global evaluation. This is not to suggest that ratings are invalid. Students’ global ratings of instructors are accurate when compared with trained observers’ ratings of specific behaviors of instructors (Murray, 1984). Thus, student ratings of instruction have a hierarchical structure, as demonstrated by Marsh (1991b), and can efficiently generate global scores of instructional effectiveness for personnel decisions.

We also agree with Marsh and Roche (1997) that validation approaches that incorporate construct validation are desirable. That is, researchers should demonstrate that the scores and interpre- tation determined by student rating forms correspond to other measures in terms of the underlying theoretical trait (Cronbach & Meehl, 1955). However, the multitrait–multimethod approach described by Marsh and Roche based on multiple raters (current students, former students, instructors themselves, colleagues, and administrators) using the same instrument has been criticized on a number of grounds (Abrami et al., 1996; Gaski, 1987). Under these conditions, the multitrait–multimethod approach can be considered to assess reliability rather than validity. Thus, we believe that it provides evidence that the different raters may share a common theory of teaching and thus rate teaching behaviors similarly.
leniency found small and inconsistent effects. In one condition, a poor-quality instructor who assigned high grades received worse evaluations from students. Thus, we are of the view that prior evidence fails to demonstrate convincingly that there are large and consistent positive biases in ratings that are attributable to grading leniency.

New evidence offered by Greenwald and Gillmore (1997) has been used by them to suggest a strong grading-leniency effect. In our opinion, evidence that student ratings are correlated with expected or actual grades does not necessarily indicate bias. Grading leniency can influence learning, and to the degree that it does so, it does not bias student ratings.

Greenwald and Gillmore (1997) constructed five alternative models to explain the grades–ratings correlation. They subsequently tested these models against their University of Washington data set. However, their methodology is appropriate only if each model is correctly specified (all relevant variables and paths are correctly specified). If this condition is not met, the exercise becomes one of setting up a straw man to knock down.

We disagree with several of the assumptions Greenwald and Gillmore (1997) made to test these alternative models (the entries in their Table 1), for example, models postulating that a third variable (e.g., true student learning causing both high ratings and high grades) could not also generate a positive within-classes grades–ratings correlation, a larger correlation with relative grades, a halo effect, or a negative between-classes grades–workload correlation (d'Apollonia & Abrami, 1997). We invite readers to create such scenarios.

Correlational methods, such as path analyses, cannot substitute for well-constructed field experiments that investigate the causal relationship between grading leniency and student ratings. Therefore, we agree with the postscript by Greenwald and Gillmore (1997) that the “best way to study the effect of grading leniency . . . is by means of experiments with grading-leniency manipulations” (p. 1217). We urge readers to consider prior experiments, such as those of Abrami et al. (1980).

Finally, the need to undertake adjustments in ratings requires that several criteria are met. First, there must be unequivocal evidence of bias. Second, simple adjustments require that the effects of the bias are uniform in size and are consistent across teaching contexts. Third, the adjustment must not affect valid influences on student ratings (e.g., teaching-quality effects). Thus, we believe that it remains to be demonstrated whether adjustments in ratings are necessary and whether adjustment schemes improve the ability of ratings to predict teaching influences on student learning and other important outcomes of instruction.

**Comment on McKeachie**

We agree with McKeachie (1997, this issue) that the multisection validity studies, by virtue of the common objective achievement test, may reflect an unsophisticated model of teaching effectiveness. In addition to McKeachie’s recommendation that student rating forms should reflect models of teaching based on cognitive and motivational research, we suggest that they should also take into consideration students’ attitudes about the evaluation process.