Validity Concerns and Usefulness of Student Ratings of Instruction

Anthony G. Greenwald
University of Washington

The validity of student rating measures of instructional quality was severely questioned in the 1970s. By the early 1980s, however, most expert opinion viewed student rating measures as valid and as worthy of widespread use. In retrospect, older discriminant-validity concerns were not so much resolved as they were displaced from research attention by accumulating evidence for convergent validity. This article introduces a Current Issues *section that gives new attention to validity concerns associated with student ratings. The section's 4 articles deal, respectively, with (a) conceptual structure (are student ratings unidimensional or multidimensional?), (b) convergent validity (how well do ratings correlate with other indicators of effective teaching?), (c) discriminant validity (are ratings influenced by factors other than teaching effectiveness?), and (d) consequential validity (are ratings used effectively in personnel development and evaluation?). Although all 4 articles favor the use of ratings, they disagree on controversial points associated with interpretation and use of ratings data.*

My interest in student ratings had a sudden onset. In 1989, I received the highest student rating evaluations I had ever received at University of Washington, for teaching an undergraduate honors seminar. The sudden interest came, not then, but a year later, when I received my lowest ever evaluations. The two ratings were separated by eight deciles according to the university's norms—about 2.5 standard deviations apart. But these two ratings were for the same course, taught in the same fashion, with a syllabus that was only slightly changed.

The two juxtaposed ratings contained more than a mild hint that my students' responses were determined by something other than the (unchanged) course characteristics or the (presumably unchanged) instructor's teaching ability. The resulting curiosity and puzzlement led to two research strategies: reading the literature and collecting data. This article describes some of what can be learned from reading the literature. This article also serves as an introduction to the following four articles by Marsh and Roche (1997, this issue), d'Apollonia and Abrami (1997, this issue), Greenwald and Gillmore (1997, this issue), and McKeachie (1997, this issue). The results of new data collections are summarized in the article by Greenwald and Gillmore.

Historical Trends in Research on Student Ratings

An electronic search for publications on student ratings (ERIC, 1966–1997; PsycINFO, 1967–1997) revealed that the topic of ratings validity has been the subject of much research, peaking in the early 1980s. Figure 1 characterizes a sample of that research in the period from 1971 to 1995. Over the entire 25-year period, more publications favored validity than invalidity. However, the research changed noticeably in character around 1980.

As can be seen in Figure 1, research activity on the validity of student ratings has declined noticeably since about 1980. The analysis of research conclusions shown in Figure 1 demonstrates that this was a specific decline in studies that remained neutral (decreasing from 31 to 16 between 1976–1980 and 1981–1985) and in studies that were critical (decreasing even more drastically, from 15 to 3). At the same time, the number of studies support-

Editor's note. Anthony G. Greenwald served as action editor for this *Current Issues* section.

The articles in this section include postscripts, in which the authors responded to an invitation to comment (in limited space) on the other articles and postscripts in this section.

Author's note. Some portions of this article were presented as an address for the Donald T. Campbell Award from the Society of Personality and Social Psychology at the 103rd Annual Convention of the American Psychological Association, August 1995, New York.

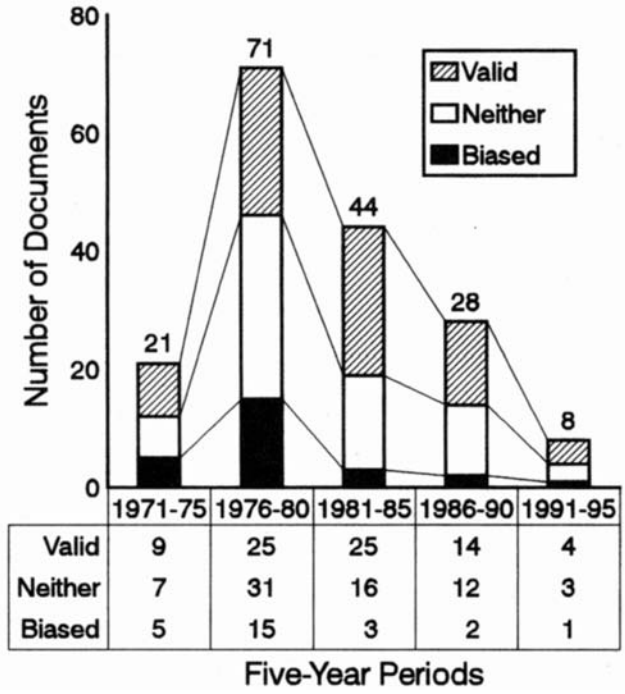
Support for this research was provided by Grant SBR-9422242 from the National Science Foundation and Grant MH 41328 from the National Institute of Mental Health.

For comments on various drafts of this article or on material preliminary to it, the author thanks Robert D. Abbott, Philip C. Abrami, Sylvia d'Apollonia, Kenneth A. Feldman, Gerald M. Gillmore, Joe Horn, George S. Howard, Herbert W. Marsh, Scott E. Maxwell, Jeremy D. Mayer, W. J. McKeachie, Robert S. Owen, Lloyd K. Stires, and John E. Stone.

Correspondence concerning this article should be addressed to Anthony G. Greenwald, Department of Psychology, University of Washington, Box 351525, Seattle, WA 98195-1525. Electronic mail may be sent via Internet to agg@u.washington.edu.

 0003-066X/97/$2.00
Vol. 52, No. 11, 1182–1186

Figure 1
Shifting Appraisals of Validity of Student Ratings



	1971-75	1976-80	1981-85	1986-90	1991-95
Valid	9	25	25	14	4
Neither	7	31	16	12	3
Biased	5	15	3	2	1

Note. This figure summarizes the author's categorization of study conclusions on the basis of abstracts retrieved from electronic searches of PsycINFO and ERIC, using for both databases the search query *(student rating$1 or teaching evaluation$1) and (bias or valid$3 or invalid$3).* The *$n* suffix included in the search any words found by appending up to *n* letters after the stem. Categorization as "biased" indicates study conclusions that student ratings of instruction are contaminated by one or more extraneous influences. The ERIC search was limited to unpublished reports in order not to have the two searches produce duplicates.

ing validity remained the same, and these increased as a proportion of the total, from a minority of 35% (25 out of 71) to a majority of 57% (25 out of 44). By the 1990s, research on the validity of student ratings had diminished to such a low level that it is easy to infer that earlier contributions had resolved the major issues.

1970s: Research Questioning the Validity of Student Ratings

Although research published in the 1970s covered a variety of concerns about validity, a major concern of that period was the possible effect of grades on ratings. The concern with grade-induced bias is apparent in the following quotes:

> The present evidence, then, supports a notion that a teacher can get a "good" rating simply by assigning "good" grades. The effect of obtained grades may bias the students' evaluation of the instructor and therefore challenges the validity of the ratings used on many college and university campuses. (Snyder & Clair, 1976, p. 81)

> The implications of the findings reported are considerable, and it is suggested that the validity of student evaluations of instructors must be questioned seriously. It is clear that . . . an instructor [who] inflates grades . . . will be much more likely to receive positive evaluations. (Worthington & Wong, 1979, p. 774)

These are conclusions from experiments in which grades had been manipulated upward or downward, and the manipulated grades were observed to raise or lower student ratings correspondingly. Several such experiments, mostly appearing in the 1970s, were conducted in actual undergraduate courses (Chacko, 1983; Holmes, 1972; Powell, 1977; Vasta & Sarmiento, 1979; Worthington & Wong, 1979). Contemporary reviews of the literature on student ratings either omit treatment of these natural classroom experiments on effects of manipulated grades on ratings or mention them only in the context of suggesting that they are collectively flawed (e.g., Abrami, Dickens, Perry, & Leventhal, 1980, p. 109; Marsh, 1987, p. 320; Marsh & Dunkin, 1992, pp. 200–202). At the same time, a meta-analytic survey of the set of classroom experiments on the effects of grade manipulation revealed that they obtained, on average, medium to large effects that were plausibly consequential in affecting actual classroom ratings (Greenwald, 1997).

Abrami et al. (1980, p. 109) speculated about possible flaws in the published experiments that demonstrated effects of manipulated grades on ratings. This speculation was stated somewhat more strongly by Marsh (1987) and even more strongly by Marsh and Dunkin (1992). Although these reservations deserve serious consideration, it must be noted also that the standard strategy for opposing published experiments on methodological grounds—repeating the experiments with improved methods—was never pursued by any of the critics of the grade-manipulation studies.[1] Consequently, the current status of the hypothesis that grading leniency–strictness affects ratings is that this hypothesis has been supported with some clarity in virtually all published experimental tests. Although the conclusions of these experiments have been questioned by critics, those conclusions have not been empirically refuted.

1980s: Demonstrations of Convergent Validity of Student Ratings

Since 1980, research on student ratings has been mostly in the form of correlational construct-validity designs. Three kinds of studies have provided evidence supporting the construct validity of student ratings.

[1] The closest to this kind of methodological follow-up was a pair of experiments by Abrami et al. (1980), which found that grades for a quiz taken after listening to a videotaped lecture had no consistent effect on evaluation of the lecturer. However, in Abrami et al.'s experiments, subjects' grades on the experimental quiz had no effect on their course grades, and the videotaped lecturer (who was not the regular teacher for the course) may not have been perceived as being responsible for the quiz grades received in the experiments.

Multisection validity studies. In the best of the largest group of construct-validity studies, multiple sections of the same course are taught by different instructors, with student ability approximately matched across sections and with all sections having identical or at least similarly difficult examinations. Using examination performance as the criterion measure of achievement, these studies have determined whether differences in achievement for students taught by different instructors are reflected in the students' ratings of the instructors. The collection of multisection validity studies has been reviewed in several meta-analyses. Although the meta-analytic reviews do not agree on all points concerning the validity of student ratings, it is clear that multisection validity studies have yielded evidence for modest convergent validity of ratings. Correlations between ratings and exam-measured achievement average about .40 (see the overview of meta-analyses by Abrami, Cohen, & d'Apollonia, 1988, especially pp. 160–162, as well as the article by d'Apollonia & Abrami, 1997).

Multisection validity studies favor construct validity of ratings by supporting an interpretation of observed correlations between grades and ratings in terms of parallel effects of a third variable, teaching effectiveness, on both measures. If grades correlate with ratings only or mainly because good teachers produce both high grades and high ratings, then all is well with the validity of student ratings.[2]

Path-analytic studies. The second type of correlational construct-validity study also explores the idea that effects of a third variable on both grades and ratings can explain their correlation but considers third variables other than teaching effectiveness. For example, Howard and Maxwell (1980) applied path-analytic techniques to show that both grades and ratings were related to measures of students' level of motivation for courses, from which they concluded that "the relationship between grades and student satisfaction might be viewed as a welcome result of important causal relationships among other variables rather than simply as evidence of contamination due to grading leniency" (p. 810). In another example of this type of study, Marsh (1980) observed that

> [a] path analysis demonstrated that students' Prior Subject Interest had the strongest impact on student ratings, and . . . also accounted for about one-third of the relationship between Expected Grades and student ratings. . . . Expected Grade was seen as a likely bias—albeit a small one—to the ratings, and even this interpretation was open to alternative interpretations. (pp. 219, 236)

Multitrait–multimethod studies. The third type of construct-validity study seeks to demonstrate that student ratings possess both convergent and discriminant validity—that is, to demonstrate that they correlate (a) relatively well with measures based on other methods for assessing the construct of quality of instruction and (b) relatively less well with measures assumed to assess other constructs (e.g., Freedman, Stumpf, & Aguanno, 1979; Howard, Conway, & Maxwell, 1985; Marsh, 1982). Such multitrait–multimethod studies typically have reported evidence for both convergent and discriminant validity of student ratings. It should be noted, however, that these researchers generally did not consider expected grades as a source of contamination.

In summary of the relatively recent literature on student ratings, and as the following quotes indicate, prominent reviews published since about 1980 give a clear impression that major questions of the 1970s about ratings validity were effectively answered and largely put to rest by subsequent research.

> In general, . . . most of the factors [that] might be expected to invalidate ratings have relatively small effects. . . . Some studies have found a tendency for teachers giving higher grades to get higher ratings. However, one might argue that in courses in which students learn more the grades should be higher and the ratings should be higher so that a correlation between average grades and ratings is not necessarily a sign of invalidity. . . . My own conclusion is that one need not worry much about grading standards within the range of normal variability. (McKeachie, 1979, pp. 390–391)

> Probably, students' evaluations of teaching effectiveness are the most thoroughly studied of all forms of personnel evaluation, and one of the best in terms of being supported by empirical research. . . . Although it is possible that a grading leniency effect may produce some bias in student ratings, support for this suggestion is weak and the size of such an effect is likely to be insubstantial in the actual use of student ratings. (Marsh, 1984, pp. 749, 741)

> [Recent] evidence has suggested . . . that rather than signaling possible contamination and invalidity of student evaluations, the observed relation between grades and student ratings might reflect expected, educationally appropriate relations. (Howard et al., 1985, p. 187)

> In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for evaluation. (Cashin, 1995, p. 6)

These quotes not only acknowledge that grades and ratings are correlated but also express the judgment that this correlation can and should be interpreted without concluding that grades create a bothersome contamination of ratings.

Contents of this *Current Issues* Section: Four Validity Concerns

The question of possible bias in ratings associated with grading leniency is a question about the discriminant validity of student ratings. Discriminant validity is just one of four types of validity with which this *Current*

[2] Interpretations of this approximate .40 correlation as reflecting processes other than, or in addition to, parallel effects of teaching effectiveness on grades and ratings also have been suggested. For example, Marsh and Dunkin (1992, pp. 173ff.) noted that some portion of this correlation could be credited either to motivational variations among students in different sections or to students' greater satisfaction with higher grades.

Table 1
Positions of the Authors of this Current Issues Section on Four Validity Issues

Authors	Validity concerns and focal questions			
	Conceptual structure: Are ratings conceptually unidimensional or multidimensional?	Convergent validity: How well are ratings measures correlated with other indicators of effective teaching?	Discriminant validity: Are ratings influenced by variables unrelated to effective teaching?	Consequential validity: Are ratings results used in a fashion that is beneficial to the educational system?
Marsh & Roche	Like effective teaching, ratings are conceptually and empirically multidimensional. Their validity and particularly their usefulness as feedback are undermined by ignoring this multidimensionality.	Different dimensions of student ratings are consistently related to effective teaching criteria with which they are most logically related, thus supporting their construct validity.	Ratings are relatively unaffected by potential biases. Bias (mis)interpretations typically fail to control valid effects on teaching (e.g., class size, enthusiasm) that ratings accurately reflect.	Multidimensional ratings, augmented by consultation, improve teaching effectiveness (their most important purpose). Their use in personnel decisions, however, should be more informed and systematic.
d'Apollonia & Abrami	Although teaching is multidimensional, ratings contain a large global factor, which consists of several highly correlated lower order factors.	Global student ratings or a weighted average of specific ratings are moderately correlated with teacher-produced student learning.	There is little evidence of bias in ratings; few characteristics have been shown to differentially affect ratings and teacher-produced student learning.	Ratings provide valid information on instructor effectiveness. However, they should not be the only source of information, nor should they be overinterpreted.
Greenwald & Gillmore	Because student ratings are dominated by a global evaluative factor, many ratings items detect only this global evaluation rather than their intended distinctive content.	Ratings measures show moderate correlations with achievement in the multisection design.	The same instructor gets higher ratings when giving higher grades or teaching smaller classes. Older research indicates also that ratings are increased by enthusiastic style.	The quest for high ratings subtly induces lenient grading, which can both (a) reduce academic content of courses and (b) feed grade inflation.
McKeachie	There is a *g* factor in ratings, but there are also discriminable lower order factors.	Student ratings provide valid, albeit imperfect, measures of teaching effectiveness.	Influences on ratings by variables other than teaching effectiveness are of concern in the context of the deplorable practice of computing ratings averages that are compared with norms.	Ratings contribute to judgments of teaching effectiveness, but their use could be improved.

Issues section is concerned.[3] The four validity questions and the positions taken on them by the authors in this section are briefly summarized in Table 1. As can be seen in Table 1, the authors of this section are of multiple minds on these four validity questions. This situation is partly by design, in that contributions to this section were invited with the explicit purpose of representing a broad range of views.

Although each of the articles in this *Current Issues* section touches on more than one validity concern, each has its strongest focus on a different one of the four validity concerns summarized in Table 1. Marsh and Roche (1997) focus on the conceptual structure of ratings. Their main concerns follow from their view of effective teaching as a multidimensional construct. Therefore, Marsh and Roche advocate the use of ratings measures that are designed to capture the breadth and the multiplicity of these dimensions. D'Apollonia and Abrami (1997) take convergent validity as their primary focus. From their overview of the multisection validity literature, d'Apollonia and Abrami conclude that ratings measures typically show substantial correlations with student achievement as measured by examination performance. Greenwald and Gillmore (1997) focus on discriminant validity, analyzing the regularly observed correlation between expected grades and evaluative ratings from multiple theoretical and statistical perspectives. Greenwald and Gillmore conclude that the strongest contributor to the grades–ratings correlation is an undesired (and statisti-

[3] Actually, two distinct discriminant-validity concerns are treated in the four articles: (a) bias in ratings (i.e., discrimination of instructional quality from other influences on ratings) and (b) multidimensionality of ratings (i.e., discrimination among components of effective teaching). The first of these is the one for which the discriminant-validity designation is used in these articles. The latter is identified as "conceptual structure of ratings."

cally correctable) influence of grading leniency–strictness on ratings. McKeachie (1997) focuses on consequential validity—how effectively ratings are put to use. McKeachie is broadly favorable on questions of convergent and discriminant validity of ratings but nevertheless is concerned that ratings are not being used effectively in many settings. McKeachie's section-concluding article both reviews the validity themes of the other three articles and surveys the complexities of using ratings for their two main purposes: evaluating teachers and improving teaching.

REFERENCES

Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research, 58,* 151–179.

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.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited* (IDEA Paper No. 32). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Chacko, T. I. (1983). Student ratings of instruction: A function of grading standards. *Educational Research Quarterly, 8*(2), 19–25.

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

ERIC [Electronic database]. (1966–1997). Palo Alto, CA: Dialog Information Services.

Freedman, R. D., Stumpf, S. A., & Aguanno, J. C. (1979). Validity of the Course–Faculty Instrument (CFI): Intrinsic and extrinsic variables. *Educational & Psychological Measurement, 39,* 153–158.

Greenwald, A. G. (1997). *Do manipulated course grades influence student ratings of instructors? A small meta-analysis.* Manuscript in preparation.

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

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

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

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

Marsh, H. W. (1980). The influence of student, course, and instructor characteristics on evaluations of university teaching. *American Educational Research Journal, 17,* 219–237.

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

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

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

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 8, pp. 143–233). New York: Agathon Press.

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

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

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52,* 1218–1225.

Powell, R. W. (1977). Grades, learning, and student evaluation of instruction. *Research in Higher Education, 7,* 193–205.

PsycINFO [Electronic database]. (1967–1997). Washington, DC: American Psychological Association.

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.

Vasta, R., & Sarmiento, R. F. (1979). Liberal grading improves evaluations but not performance. *Journal of Educational Psychology, 71,* 207–211.

Worthington, A. G., & Wong, P. T. P. (1979). Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology, 71,* 764–775.

Postscript

D'Apollonia and Abrami (1997) as well as Marsh and Roche (1997) responded to this article's brief mention of a meta-analysis (Greenwald, 1997) of a small set of natural classroom experiments that were conducted in the 1970s and used grading strictness–leniency manipulations. The concerns of d'Apollonia and Abrami as well as Marsh and Roche are that this meta-analysis was too uncritical of the set of older experiments. My concerns are just the opposite—that (a) published critical reviews have led to insufficient attention to these older experiments, and (b) the published negative evaluations of these experiments may have discouraged conduct or publication of subsequent similar studies. Regardless of one's opinion of these older experiments, it is clear that additional experiments in natural classroom settings could help to resolve uncertainty about causal effects of grading strictness–leniency manipulations.