Research in Higher Education, Vol. 30, No. 2, 1989

HOW SHOULD WE USE STUDENT RATINGS TO EVALUATE TEACHING?

Philip C. Abrami

The proper use of student ratings to evaluate teaching is the subject matter of this paper. In particular, two divergent views are contrasted: the multidimensional view of evaluation and the unidimensional. Marsh (*Int. J. Ed. Res.*, 1987, v. 11, pp. 253–388) is the most outspoken advocate of the multidimensional view of evaluation based on factor scores. Because we disagree over such an important use of ratings, I explain the position I have adopted, summarize Marsh's most recent position, and respond to his concerns.

There are now very few universities in North America at which the evaluation of teaching effectiveness is undertaken without recourse to student ratings of instruction. The increasingly widespread use and acceptance of student ratings has not been accompanied by uniform evaluation procedures. There are dozens, perhaps hundreds, of instruments in use ranging from those generated by sophisticated, commercial item banks to locally developed forms with only a handful of items. Furthermore, the information these forms provided is used in fundamentally different ways to judge teaching effectiveness.

The proper use of student ratings to evaluate teaching is the subject matter of this paper. In particular, two divergent views are contrasted: the multidimensional view of evaluation and the unidimensional. Each view has found support among both evaluation experts and users. For example, Johnson (1989) surveyed experts and found that an equal number favored as were opposed to the use of only global ratings for promotion and tenure decisions. Yet only one of these views is best to evaluate teaching.

Marsh (1984, 1985, 1987) is the most outspoken and articulate advocate of the multidimensional view of evaluation. He suggests that because teaching is

An earlier version of this paper was presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April, 1988.

Philip C. Abrami, Education Department, Concordia University, Montreal, Quebec, Canada H3G 1M8.

0361-0365/89/0400-0221$06.00/0

multifaceted, student ratings should not be summarized by a response to a single item or an unweighted average response to many items. Instead, evaluations of teaching for summative or formative purposes should be based on factor scores from instruments such as his Student's Evaluations of Educational Quality (SEEQ).

Along with others, I disagree with many aspects of this suggestion, particularly when ratings are used for promotion and tenure decisions. I also have reservations about the use of certain factor scores when ratings are used for instructional improvement. Because we disagree over such an important use of ratings, I will further explain the position I adopted elsewhere (Abrami, 1985, 1988, 1989), summarize Marsh's most recent position, and respond to his concerns.

CONCERNS ABOUT FACTOR SCORES

My concerns with the reporting and use of separate factor scores for summative purposes are multiple. They reflect several problems which further research may solve, but to date has not solved sufficiently to justify the use of separate factor scores in promotion and tenure decisions. Consequently, for summative purposes, I favor the use of several global rating items (e.g., How would you rate this instructor in overall ability?) or a carefully weighted average of rating factors in lieu of separate factor scores.

First, I do not believe we have sufficient evidence to establish either what the dimensions of effective teaching are or whether and how they are interrelated. There have been many factor analyses of student rating forms, but they do not lead to identical descriptions of effective teaching. The lack of good theories about teaching and the inconsistent results of factor analyses makes me question any one operational definition of the facets of teaching.

While I am certain that teaching is almost always multidimensional I am not convinced that the nine dimensions of the SEEQ represent those dimensions invariantly for all instructors, courses, students, and settings. Nor am I convinced that the different characteristics measured in other carefully developed, well-validated ratings forms or the nineteen dimensions described in a review of forms (Feldman, 1976) are the characteristics of teaching.

I discussed the inconsistent findings and methodological problems with prior factor analytic work in a previous paper (Abrami, 1985). In more recent work, Abrami and d'Apollonia (1989) attempted to categorize, using Feldman's dimensions, the findings from forty-three studies of the validity of student ratings in multisection college courses. Wherever possible, we located the many rating forms employed in the studies so that we could categorize the findings item-by-item. We found factors in some studies whose items fit into many

dimensions and factors from other studies whose items fit into only a single dimension. We also found dimensions heavily represented in some rating forms and absent in others.

Second, I have concerns about the *content* validity of specific items and some of the dimensions they compromise when ratings are used across a wide variety of courses, instructors, students, and settings. For example, the appropriateness of items on Rapport and Interaction is different in large classes than small, in discussion (or studio) classes than lectures, etc.

For a multidimensional rating form to have content validity, its items must elicit a representative sample of student descriptions from the relevant domains of instructor behavior. A rating form should not contain too many items assessing one sort of instructor behavior and too few items assessing another. In addition, a rating form should contain items equally relevant to each of the instructional situations for which it was designed.

Of what relevance are the SEEQ items "Students were encouraged to participate in class discussion" and "Instructor was friendly toward individual students" in classes with very high enrollments, especially when compared to small classes? Imagine the instructor who encouraged each student to participate when class size was over one hundred. The instructional situations are so different that it is wrong to suggest that these instructional characteristics are equally relevant. Yet this is the presumption which underlies every universal, multidimensional rating form where these factors appear.

Third, Cohen's (1981) quantitative review of the multisection validity studies suggests that many rating dimensions have lower correlations with student learning (e.g., Rapport = .31, Interaction = .22, Feedback = .31, Evaluation = .23) or *near zero* correlations with student learning (Difficulty = −.02) compared with Overall Course (.47) and Overall Instructor (.43) correlations with learning. Thus, the *construct* validity of some rating factors is poor as it can be inferred from research on ratings and teacher-produced student achievement.

Fourth, we know much less about the *generalizability* of specific rating factors than global ratings. Specifically, we know less about how well rating dimensions are construct valid (measure effective teaching) under a variety of course, instructor, student, and setting conditions than global ratings. Furthermore, we can expect the validity of the specific factors to vary across situations as the factors are often reasonably independent of each other. Knowing that one rating factor is uninfluenced by a biasing characteristic gives little assurance about the absence of the bias in other factors.

Fifth, we cannot expect administrators or nonexperts in evaluation to properly weigh the information provided by factor scores in arriving at a single decision about the quality of an instructor's teaching. This is particularly troublesome when comparative judgments about teaching are made. We cannot expect

administrators to have the expertise of instructional evaluators, nor have we provided them with precise and defensible procedures for synthesizing the information from factor scores. My experience is that administrators weigh factor scores equally or look for particularly strong or weak areas of teaching. Personally, I would be disappointed to learn that a faculty member was denied tenure because of low student ratings on "Difficulty" when such ratings correlate near zero with student learning.

MARSH'S CONCERNS ABOUT GLOBAL RATINGS

Marsh favors a multidimensional approach to student ratings, even for summative purposes. Student ratings should be multidimensional because teaching is multidimensional. If a rating form contains an ill-defined collection of items and ratings are summarized by an average of those items, there is no basis for knowing what is being measured and no basis for weighting components in a way that is appropriate to the particular purpose to be served. Marsh describes five reasons why factor scores are preferable to a total rating or an overall rating:

1. There are many possible indicators of effective teaching; the component that is most valid will depend on the criteria being considered.
2. Reviews of different validity criteria show that components of ratings are more highly correlated than an overall or total rating.
3. The influence of biasing characteristics is more difficult to interpret with total ratings than with specific components.
4. The usefulness of ratings for formative purposes is enhanced by the presentation of factor scores.
5. Even if it were agreed that ratings should be summarized by a single score for a particular purpose, the weighting of the factors should be a function of logical and empirical analyses.

   Marsh adds two additional criticisms of total or overall ratings directed at researchers who accept that ratings are multidimensional but argue that personnel decisions are unidimensional:

6. There appears to be no empirical research to support the claim that administrators are unable to utilize or prefer not to be given multiple sources of information in their deliberations.
7. The use to which ratings are put has nothing to do with their dimensionality. It may, however, influence the form in which the dimensions are presented.

RESPONSE TO MARSH'S CONCERNS

I will address Marsh's concerns point for point. Although in the end we recommend very different rating scores, we do agree on several issues.

First, there are many possible indicators of effective teaching, but unfortunately most of the direct products of instruction are not articulated in any theory of effective teaching, and are not well operationalized. There is little good research to establish the complex network of relationships between student impressions of the processes of instruction and the impacts of those processes on student cognition and affect.

For example, most studies of the relationship between ratings and teacher-produced student learning have dealt with learning at the lowest level of the Bloom taxonomy. If higher-level learning was carefully studied, the dimensions of effective instruction might change. Furthermore, indirect measures of teaching effectiveness (e.g., peer review, self-evaluation) are questionable indices on psychometric and conceptual grounds.

Therefore, I agree with Marsh that ratings are multidimensional and probably should relate to effectiveness criteria differently. The question is how should they relate and how do we then weight these relationships, particularly when arriving at a single decision about the quality of an instructor's teaching.

Second, Cohen's (1981) review of multisection validity studies revealed that only one dimensional rating, Skill (which is a fairly general rating factor), correlated more highly with teacher-produced student learning (albeit lower-level) than global ratings. That correlation (.50) is uselessly greater than those for global ratings. In contrast, some of the remaining dimensional ratings have meaningfully lower correlations than global ratings. In terms of the single most important product of instruction—student learning—Marsh is wrong to assert that dimensional ratings are more highly correlated with the validity criterion than global ratings.

Third, while the effect of biasing characteristics may be more difficult to understand when ratings are global than multidimensional, the effects of biases on multidimensional ratings are no less severe than those which affect unidimensional ratings. For example, it may be easier to understand why instructor expressiveness affects student ratings of "Rapport" than it is to understand the effect of expressiveness on global ratings. In both instances, however, expressiveness affects ratings more than student learning and thus represents a bias.

Furthermore, multidimensional ratings introduce additional and more complex biases than global ratings; some rating dimensions are sensitive to biasing effects not present in global ratings. Thus for practical uses, greater care must be exercised to control particular biases for particular rating dimensions than is necessary for global ratings.

Fourth, I don't think that global ratings are useful for formative purposes except perhaps to warn faculty that change is necessary and to motivate them to improve their teaching. And I agree with Marsh that dimensional ratings need to

be validated even when the purpose of evaluation is "only" instructional improvement.

I have always felt somewhat uncomfortable with cafeteria approaches to instructional evaluation (e.g., where a limited selection is made from a large bank of specific, behaviorally oriented items) because of the tenuous assumption that instructors must know the qualities of good teaching for themselves, their courses, and their students as reflected in their selection of items for rating. In the cafeteria approach instructors learn whether the *delivery* of these characteristics needs improvement through the item ratings which are subsequently received. However, they do not learn whether the characteristics they have selected are the appropriate ones (i.e., affect student learning) for that situation.

Nevertheless, the cafeteria approach recognizes that most multidimensional ratings scales are developed with general use in mind, which limits their application to particular subject matters and particular classes. Thus, the item bank approach is a recognition that the analysis of instructional effectiveness provided by a fixed set of factors made up of nonspecific items is inadequate in any single situation and reduces the usefulness of student evaluation for instructional improvement.

Fifth, Marsh and I agree that a weighted average of factor scores is superior to an unweighted average. That would generally mean a zero weight for Difficulty ratings. I wonder how many authors of multidimensional rating scales are prepared to accept that kind of weighted average?

Sixth, there is as yet no empirical research to suggest that administrators are unable to utilize or prefer not to be given multidimensional ratings. For now, common sense and practical experience should suffice to tell us that administrators are neither as well informed as researchers nor certain to weigh rating factors in proportion to their validity. It's better to provide them with a weighted average or a global rating.

Seventh, the use to which ratings are put has everything to do with how one perceives their dimensionality. In an earlier paper (Abrami, 1985), I argued that ratings could be both multidimensional and unidimensional in a fashion analogous to the way Wechsler conceived of intelligence as composed of specific components which combined to form some general intellectual ability.

I have another analogy to convince the doubters, and ironically this one is provided by L.L. Thurstone, the father of factor analysis. When Thurstone (Thurstone and Chave, 1929) developed procedures for the measurement of attitudes, he described how it was possible to measure attitudes toward a topic such as religion and the church, which clearly had so many different dimensions, along a single, unitary dimension. Thurstone explained that multidimensional concepts could be evaluated on a unitary dimension in much the same way as we organize our perceptions of physical objects. So a piece of

wood can have a particular, height, width, length, weight, color, etc., but we have little trouble recognizing that wood as a table.

So too with good teaching. It does make conceptual and empirical sense to speak of effective teaching as a unidimensional concept and to make summative decisions about teaching using a unidimensional rating. This choice then frees us to recognize that the particular characteristics of effective teaching vary across instructors, courses, students, and settings—that the specific dimensions of teaching should and do vary.

Acknowledgments. Preparation of this article was aided by a grant to the author from the Ministry of Education, Providence of Quebec, Canada. The author expresses his appreciation to Wilbert J. McKeachie for his helpful comments on an earlier draft of this paper.

REFERENCES

Abrami, P. C. (1985). Dimensions of effective college instruction. *Review of Higher Education* 8: 211–228.

Abrami, P. C. (1988). SEEQ and ye shall find: A review of Marsh's "Students' evaluation of university teaching". *Instructional Evaluation* 9(2): 19–27.

Abrami, P. C. (1989). SEEQing the truth about student ratings of instruction. *Educational Researcher,* 18(1): 43–45.

Abrami, P. C., and d'Apollonia, S. (1989). The validity of student ratings of instruction: A final review (for now). Manuscript submitted for publication.

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

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

Johnson, G. R. (1989). Faculty evaluation: Do experts agree? *The Journal of Staff, Program, & Organization Development,* 7:22–28.

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

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

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research* 11: 253–388.

Thurstone, L. L., and Chave, E. J. (1929). *The Measurement of Attitudes.* Chicago: University of Chicago Press.

Received January 4, 1989