10. The Dimensionality of Student Ratings of Instruction: What We Know and What We Do Not*

Philip C. Abrami[†], Sylvia d'Apollonia[‡] and Steven Rosenfield[§]
[†]*Concordia University*
abrami@education.concordia.ca
[§]*Vanier College*

Key Words: Student ratings, effective teaching, post-secondary education, meta-analysis

Sometime during the second half of almost all college and university courses offered in North America, a brief ritual occurs. Students take out their sharpened pencils (number two lead, if you please) and quickly answer a series of multiple choice questions covering a range of issues about the course and their instructor. Student rating forms often contain specific items, which are purported to reflect a number of distinct dimensions of instructional effectiveness, as well as a few global items, which reflect students' overall impressions of the instructor and the course. Examples of specific items include: "Does the instructor have a good command of the subject matter?" "Does the instructor use class time well?" "Is the instructor friendly?" "Does

*An earlier version of this paper was presented at the annual meeting of the American Educational Research Association, Atlanta, Georgia, April, 1993 as part of a symposium entitled: "Student ratings of instruction: Meta-analysis of their dimensionality."

This research 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).

The authors gratefully acknowledge the constructive feedback received from Harris Cooper, Kenneth Feldman, Wilbert McKeachie, Herbert Marsh, Raymond Perry, and an anonymous reviewer who read earlier versions of this manuscript.

Address reprint requests to: Dr. Philip C. Abrami, Centre for the Study of Classroom Processes, Concordia University, 1455 DeMaisonneuve Blvd. W., Montreal, Quebec CANADA H3G 1M8.

R.P. Perry and J.C. Smart (eds.), The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective, 385–456.

the instructor assign difficult reading?" "Does the instructor facilitate class discussion?" "Does the instructor keep students informed of their progress?" Examples of global items include: "How would you rate the instructor in overall ability?" "How would you rate the quality of this course?" "How much have you learned in this course compared to others?" Many student rating forms also provide students with the opportunity to provide narrative feedback about the course, the instructor, and their learning. While the rating ritual ends quickly, the implications of the results can be far reaching, for student ratings are used for a variety of important purposes.

In many circumstances ratings are the most influential or only source of information on teaching available for decisions about promotion, tenure, or merit. Typically, personnel committees use ratings to judge teaching effectiveness by comparing individual faculty results with departmental norms. Ratings are also widely used for instructional improvement to provide feedback to instructors on the quality of their courses. Faculty use ratings feedback to identify both areas of strength that should be maintained and areas of weakness that require modification. Ratings are occasionally used by students as a guide to course selection. For example, some students may use ratings information to select the highest rated instructors, while others may use ratings information to select the easiest courses. Thus, student ratings serve widespread and important practical purposes.

Student ratings also serve important theoretical purposes by providing researchers with information on the teaching-learning process. For example, such information may be useful in assessing the effectiveness of innovative pedagogical techniques such as cooperative learning, in understanding the relationship between instructional preparation and delivery as they affect multiple outcomes of instruction, and in judging the impact of instructional strategies for different students, courses, and settings.

The practical and theoretical utility of student ratings depends on the extent to which ratings meet psychometric standards of excellence. Concerns about the reliability, validity, and generalizability of student ratings include: Are rating results consistent over time? Are students uniform in their assessments of instructors? Are ratings free from the influence of biasing characteristics? What is the dimensionality of student ratings? Are these dimensions consistent across students, courses, settings, and rating forms? Which dimensions reflect the impact of instruction on student learning and other outcomes?

This paper is concerned with the dimensionality of instruction as reflected in student ratings. Research on the dimensions of effective teaching is not new. There are numerous studies which have explored this issue and notable disagreements (e.g., Abrami, d'Apollonia and Cohen, 1990; Marsh, 1987) regarding, in particular, whether and how data from multidimensional student rating forms should be used in summative decisions about teaching (e.g., promotion, merit, tenure, etc.). This paper critically examines many of these issues and reaches important conclusions about the dimensionality of teaching as reflected in student ratings, makes practical suggestions, as well as suggests directions for future research.

In the first section, three alternative definitions of effective teaching are presented and critically analyzed: the product definition, the process definition, and the process-product definition. We contend that the relationships between teaching processes and teaching products is of major interest to researchers and practitioners.

The second section provides a general discussion of methods for empirically determining effective teaching with special emphasis on the use of student ratings for each of the three definitions of effective teaching. We comment on the difficulties of directly assessing the products of instruction and suggest the use of a table of specifications as one way to develop a rating form to indirectly measure what and how students have learned. We suggest that student ratings as process measures must contain items which assess the relevant aspects of teaching accurately in each instructional context. We note that the dimensionality of student ratings varies with course characteristics and we suggest that some items which evaluate specific aspects of teaching vary in relevance across contexts. We show that multidimensional student rating forms do not contain items which evaluate the same, specific teaching qualities; the rating forms lack both comprehensiveness and uniformity. We conclude that since the qualities of teaching evaluated by different student rating forms appear to differ both in their nature and structure, it is of value to explore the forms further and determine if there are dimensions of teaching common to a collection of student rating forms.

The third section concentrates on the strengths and weaknesses of three validation designs—the laboratory design, the multisection validation design and the multitrait-multimethod design—for empirically determining the relationship between the processes and products of teaching. The laboratory design uses the experimental manipulation of instructional conditions to study the causal effects of instruction on students. It is often considered low in external validity. The multisection validation design uses multiple sections of the same course taught by different instructors employing common measures of student ratings and student learning. The correlations between course section means for student ratings and means for student achievement explore the relationship between instructional processes and an important instructional product. We consider the multisection design particularly strong because it reduces the probability of rival explanations to instructor impacts and is high in generalizability to classrooms. In the multitrait-multimethod design, student ratings and several criterion measures (e.g., instructor self-ratings) are collected across a wide range of courses, without controlling for biasing or extraneous influences. We consider this design weaker both in internal validity, since controls are lacking, and in external validity, since important product measures of instruction (e.g., student learning) are not included. We conclude that studies employing the multisection design are worthy of special attention.

The fourth section examines the quantitative reviews of the 43 multisection validity studies. We describe what we have learned from these studies and what remains to be learned of the relationship between what instructors do when they teach and how this affects student learning. We note that reviews to date suggest that the specific dimensions of teaching appear to differentially and, in some cases, poorly predict instructor impacts on learning compared to global ratings. We suggest that there are several limitations of prior reviews. First, the reviews include only a fraction of the findings from the original studies. Second, there is the lack of a comprehensive, empirically validated system for organizing the findings from different rating forms into a common framework. Third, study features which may explain the variability in study findings remain unexplored. Consequently, a more comprehensive research integration is called for using an empirically determined scheme for coding the findings from different rating forms.

The fifth section summarizes our attempt to identify the common dimensions of effective teaching as reflected in student ratings. First, we summarize our reanalysis of Marsh in which we failed to find many specific teaching dimensions but found a general teaching factor instead. Since our ultimate goal is to explore the relationship between process and product, we concentrate on the rating forms used in the 43 multisection validity studies. We quantitatively integrate the results from 17 inter-item correlation matrices by: a) coding the items using a common scoring scheme, b) eliminating items which were heterogeneous within categories, and c) factor analyzing the aggregate correlation matrix.

Our factor analysis indicates that there is a common structure to instruction. Four factors emerged of which the largest ones were highly correlated. We conclude that existing analyses provide support for a large underlying general trait although it may not be the only trait. We also believe that effective teaching is multidimensional but that there are differences across rating forms concerning the specific dimensions which underlie effective instruction. These differences suggest that student ratings of specific teaching dimensions should not be used indiscriminately for summative decisions about teaching effectiveness. Now that we have identified the common structure of student ratings, the next phase of research will be to use the techniques of quantitative research integration to explore the relationship between this structure and teacher-produced student achievement as well as the substantive and methodological variables which explain inconsistencies in the relationships.

DEFINITIONS OF EFFECTIVE TEACHING

Effective teaching can be defined from several perspectives. In the first perspective, effective teaching is defined in terms of affecting student products. In the second perspective, effective teaching is defined in terms of the processes which instructors enact. These views are elaborated and contrasted below. The relationship between process and product views is also presented. The relationship between the process and product views of effective teaching seeks to find the links between what teachers do and whether and how students change as a result.

THE PRODUCT DEFINITION OF EFFECTIVE TEACHING

Broadly speaking, effective teaching from the product view can be defined as the positive changes produced in students in relevant academic domains including the cognitive, affective, and occasionally the psychomotor ones (to use the general taxonomic classifications developed by Bloom et al., 1956). Included in the cognitive domain are both specific cognitive skills (e.g., subject matter expertise), general cognitive skills (e.g., analytical thinking), and meta-cognitive skills (e.g., error correction). Included in the affective domain are attitudes and interests toward the subject matter in particular and learning in general as well as interpersonal skills and abilities relevant to learning and working in a social context. Finally, included in the psychomotor domain are physical skills and abilities ranging from those acquired in a physical education to precise motor skills acquired in a fine arts education.

This definition concentrates on the products that effective teaching promotes in students. The definition has several corollaries. First, there is not a single product of effective teaching; there are many. Second, there is no *a priori* theoretical requirement that the products are interrelated either within or across domains. For example, it is not necessarily the case that increased student knowledge of basic facts will result in increased analytical and synthesis skills or vice versa. Third, the value attached to individual products is often situation-specific, requiring adjustments to meet the local needs described by students, departments, and colleges. Fourth, greater teaching effectiveness is not necessarily associated with the number of products affected. Fifth, the definition makes no prediction about the (casual) sequences or paths among products. For example, it does not explicate whether student casual beliefs about learning affect academic self-concept or vice versa.

The product definition of effective teaching recognizes that there is widespread disagreement in the academic community about both the objectives and goals of instruction and the ways to achieve them. For example in the social sciences, clinical practitioners may dispute experimental researchers about the importance of developing the affective skills of students. While almost all faculty will agree with the preeminence of developing the cognitive abilities of students, there is less general agreement over the form that development takes. For example, in the natural sciences physicists may dispute whether to teach about the many concepts of the discipline or how to teach students to discover a few fundamentals.

The Process Definition of Effective Teaching

The process definition of effective teaching emphasizes the acts of teaching rather than the consequences of those actions. The process definition is meant to include instructor activities which occur both before (preparatory) and during (delivery) teaching. Preparation may include such wide-ranging activities as: developing content expertise; preparing course outlines, activities, and objectives; selecting a teaching method; assigning course workload; and setting evaluation practices and procedures. The delivery procedures may include classroom activities and abilities such as organization, dynamism, enthusiasm, and rapport, and outside classroom activities such as availability to, and friendliness toward, students.

The process definition has several corollaries. First, there is not a single process of effective teaching; there are many. The definition recognizes that effective teaching is multidimensional consisting of numerous and apparently distinct acts. Second, the definition is tentative regarding the specific acts which constitute the process. One purpose of our research is to determine empirically whether there is uniformity and consistency to these acts. Third, it is also possible that these distinct acts represent different operationalizations of an underlying construct or constructs. For example, "instructor clarity" may consist of clarity of speech, audibility, pace, comprehensibility, etc. Furthermore, these constructs may be both additive and hierarchical. This is also an empirical question. Fifth, the term "effective teaching" means that there is an evaluative component to the process. This evaluative component regards both the instructor's choice of acts and the quality and quantity with which they are enacted. In other words, ineffective instructors may emphasize the wrong acts when they teach or enact them poorly.

It is also unclear whether generally static personal characteristics or traits (e.g., gender, race, age, personality, etc.) form part of the process definition. They are qualities which are beyond the control of the instructor but which may nevertheless indirectly influence both the acts of teaching and the products of teaching. These are sometimes referred to as biasing characteristics in recognition both of their potential for influence and the undesirability of that influence.

THE PROCESS-PRODUCT DEFINITION OF EFFECTIVE TEACHING

What activities differentiate good instructors from poor ones in promoting students' critical thinking, task engagement, and persistence? Is instructor enthusiasm an important teaching process because enthusiasm motivates students to learn? Important questions such as these speak to the inexorable link between teaching processes and products.

It is our contention that the relationships between teaching processes and teaching products is of major interest. The link between process and product raises new questions about the meaning of the term "effective teaching." Now, rather than effective teaching being defined only in terms of either process or product, we may combine the two. Doing so helps identify links between what teachers do and whether and how students change as a result.

Broadly speaking, effective teaching from the process-product view can be defined as the instructor activities which occur both before (preparatory) and during (delivery) teaching which produce positive changes in students in relevant academic domains including the cognitive, affective, and occasionally the psychomotor ones.

We hypothesize that the varied products of effective teaching are affected by different teaching processes. But we cannot describe with any great confidence the specific nature of these causal relationships.

We further hypothesize that the causal relationship between any one teaching process and any one teaching product will vary as a function of external influences including student, course, and setting influences. As stated previously, there appears to be important disagreements among faculty on what to teach and how to teach it.

To summarize, we have briefly explored three alternative definitions of effective teaching: the product definition, the process definition, and the process-product definition. We believe the relationship between teaching processes and teaching products is of major interest.

EMPIRICALLY DETERMINING EFFECTIVE TEACHING

In this section we consider ways to determine effective teaching empirically for the three definitions of teaching presented. We concentrate, in particular, on the use of student ratings for these purposes.

Empirically Determining the Products of Effective Teaching

According to the product definition, effective teaching produces changes in such student outcomes as content knowledge, analytic ability, academic self-concept, motivation to learn, aesthetic appreciation, and so on. Unfortunately, the authors are unaware of individual studies that attempt to systematically and inclusively describe college teaching from a product-based perspective. There are studies that explore outcomes singly, particularly those that examine the effects of teaching on (undifferentiated) student learning of course content. Therefore, it may be profitable to apply the techniques of quantitative research integration to the literature on instructional products to better and more completely understand the effects of teaching.

In recent years, Seldin (1991), Shore et al. (1986), and others have argued for the use of the teaching portfolio, a comprehensive collection of descriptive and evaluative information on individual faculty teaching, which might include a statement of teaching responsibilities, course syllabi, instructor self evaluations, a description of improvement efforts, peer assessments, participation in teaching conferences, videotapes of instruction, student exams and essays, alumni ratings, and so on. The portfolio is to be used both for teaching improvement purposes and for summative decisions.

Judging teaching effectiveness by examining the evidence of student accomplishments—tests, papers, and projects—generally requires that two criteria are met: a) the data presented are representative of the faculty member's effect on students and b) the results of faculty can be objectively compared. Meeting the first criterion requires examining the results either of all students or a random sample of students. Submitting the best student products as evidence of teaching effectiveness, a common practice, does little to allow accurate judgments of how well instructors promote student learning.

Meeting the second criterion requires measures of student productivity that can be compared across courses. Unfortunately, this has rarely been accomplished. For example, it is extremely difficult to compare the achievement of students enrolled in an introductory Physics course with the achievement of students enrolled in an advanced, upper-level Physics course in order to judge which instructor best promotes student learning. Are differences in achievement between the two courses due to the quality of the students enrolled? The difficulty of the tests used? The nature of the material learned? The quality of the instruction given? Similarly, it is tenuous to assume that changes from pretest examination scores at the beginning of term to posttest examination scores at the end reflect only the impacts of instruction. In contrast, it is less difficult to compare student achievement on a final, common examination when the students are enrolled in different sections of the same course, especially when it is reasonable to assume that students selected course sections more or less at random. Under circumstances resembling the latter, using product measures to compare and judge instruction seems quite defensible and its use should be more widespread. In general, however, product measures of effective teaching are seldom practical to use and rarely provide accurate data for judging quality teaching.

*Student ratings as **direct** product measures*. Student ratings measure directly one product of instruction; namely, student satisfaction with teaching. For many, measuring student satisfaction with teaching is a sufficient reason to use student ratings. Proponents of the use of ratings as satisfaction measures argue that if students are the consumers of the teaching process, then student satisfaction with teaching should be a component of instructional evaluation.

Otherwise, student ratings do not measure *directly* how much or how well a class of students has learned or any other aspect of achievement in the cognitive domain including how well the content is retained. Student ratings also do not often measure directly: most affective products of instruction such as student expectations, beliefs, and concepts about themselves as learners; student attitudes, values, and interests toward the subject matter including enrolling in other courses in the area or adopting the area as a field of major study; student interpersonal and social skills generally and such skills within the context of executing a complex academic task; etc.

Student ratings as ***indirect*** *product measures.* Student ratings are often used as convenient alternative measures of most instructional products. Ratings are used to *infer* that highly rated instructors positively affect instructional products. Student ratings provide a basic yardstick for these judgments when product measures are unavailable, when the product measures are of questionable quality, or when conditions (such as differences in the level or type of course) do not allow for fair comparisons of products across instructors.

To what extent do student ratings reflect the impact of instructors on students learning of course content, their motivation to learn, development of interpersonal skills, and so on? There is a reasonable body of well-designed research, reviewed more extensively elsewhere in this paper, which suggests that, on average, there is a modest, positive relationship between global ratings of instruction and instructor-produced student learning of lower-level academic skills (e.g., knowledge of basic facts, simple comprehension, etc.). Much less is known about the validity of ratings as predictors of other outcomes of instruction.

Improving student ratings as ***indirect*** *product measures.* Consider the following item from a student rating form: "Rate the extent to which your instructor motivated you to learn." Does this item ask students to describe an instructional process or an instructional product? The item does not ask students to judge instructor preparation or delivery but the consequences of teaching. It is, therefore, not a measure of a teaching process. But is it is an accurate, *indirect* assessment of an instructional product? It is accurate only to the extent that student self-report of motivation reflects student persistence at learning, the intensity of student effort, student choice of tasks to learn, etc. Rating forms occasionally include items that ask students to assess the success of instructors at encouraging them to learn but seldom include items that assess the specific behaviors associated with that motivation.

Table 1: Table of Specifications for Student Ratings of Course Content in Psychological Statistics

Instructions: Please use this rating form to assess how well your instructor taught you the content of this course. Begin by assigning your instructor an overall rating for the amount you learned in the course. Use the box with the darkest shading for this purpose. The major content areas of the course are listed in the *rows* of the table. For each content area or row assign your instructor an *overall* rating using the scale shown below. For example, if your instructor taught you descriptive statistics extremely well assign an overall rating of 5 for descriptive statistics. The major cognitive objectives of the course are listed as *columns* in the table. For each cognitive objective or column assign your instructor an *overall* rating. For example, if your instructor taught you to apply the content extremely well assign an overall rating of 5 for application. Finally, use each box to give your instructor a rating for both what you learned and how you learned it. For example, assign your instructor a "4" if (s)he did a very good job teaching you to evaluate uses of the t-test.

Use the following rating scale in making your judgments:
1—Poor
2—Fair
3—Good
4—Very good
5—Excellent
NA—*Not applicable*

	How the Content Was Learned					
Course Content	Knowledge	Compre-hension	Appli-cation	Analysis	Synthesis	Evaluation	OVERALL RATING
Descriptive statistics							
The t-test							
Oneway Anova							
Factorial Anova							
Nonparametrics							
OVERALL RATING							

Similarly, rating forms do not often contain items that ask students to assess an instructor's impact on specific cognitive and meta-cognitive achievements. Instead, rating forms more frequently ask students to rate: "How much have you learned in this course compared with others?" A questionnaire can be designed so that ratings items may be made more precise by asking students to judge how well they learned from the instructor in each content area of the course as well as the depth to which they learned. (See Table 1, page 222.)

The table of specifications or teaching blueprint presented in Table 1 illustrates a student rating form for an undergraduate course in psychological statistics. The rows represent the content to be learned.

The columns represent how the content is to be learned. The cells or boxes represent the combination of what is to be learned and how it is to be learned. Students may use this type of evaluation form to judge an instructor's effectiveness: overall in promoting student learning, in particular content areas of the course, and in promoting different types and levels of learning. The evaluation form also allows for very specific feedback on particular aspects of teaching. For example, was the instructor effective at promoting higher level skills in more complex areas of the course?

Not all the content areas of the course are equally important, nor is every type of learning of equal value and emphasis. For example, the instructor may need to spend considerable time on some topics (e.g., descriptive statistics) and not others (e.g., factorial ANOVA). Similarly, some topics may require substantial efforts devoted to basic knowledge and comprehension while other topics may require greater efforts devoted to analysis, synthesis, and evaluation.

Prior to the evaluation, the instructor and/or students may wish to estimate the amount of time devoted to each content area and type of learning. First, estimate the percent of course time devoted to each content area. The sum of the row percentages should be 100%. Next estimate the percent of course time devoted to each cognitive objective. The sum of the column percentages should be 100%. Next, fill in each cell or box percentage. Note that the precision of the table of specifications rating form in assessing student learning remains to be determined empirically.

Finally, not all rating items seem as logically defensible as indirect measures of instructional products as the self report items described above. For example why should instructor friendliness and openness toward students necessarily reflect student understanding of thermodynamics? Indeed, we recall rather heated discussions by some faculty that they do not. Therefore, such items are better understood to reflect student ratings of the processes of effective teaching. An interest in whether such items and similar items can be used to assess instructor impacts on student learning and other outcomes is, consequently, an interest in the relationship between process and product.

Empirically Determining the Processes of Effective Teaching

Many studies have attempted to determine empirically the dimensions, clusters, factors or major characteristics that college instructors employ. Are these characteristics too many or too varied to describe succinctly?

Do faculty and students agree on the characteristics they describe? Can these characteristics be grouped together?

A major portion of the research has relied on empirical methods for identifying teaching dimensions, chiefly through the use of factor analysis. Marsh (1987) summarized research on one instrument, Students' Evaluations of Educational Quality (SEEQ), which identified nine factors of instruction. Marsh (1987) argued for consideration of these nine factors when summative evaluations of teaching are made (e.g., for promotion and tenure decisions).

Feldman (1976) reviewed studies in which students were asked to describe the characteristics of best teachers, or of ideal teachers, or of good teaching. He identified 19 dimensions which he used to classify the descriptions. Later, Feldman (1988) reviewed studies comparing faculty and student specifications of the instructional characteristics they considered particularly important to good teaching and effective instruction. In the latter review, Feldman (1988) identified 22 instructional dimensions. The average correlation between students and faculty in their judgement of these components was +0.71. Feldman (1988) concluded that there was general agreement between faculty and students in their views of good teaching as reflected in the importance the two groups placed on the components of teaching.

Feldman (1976) and Kulik and McKeachie (1975) reviewed factor analytic research of student ratings of instruction. Feldman (1976) employed 19 categories to categorize the items from 60 studies. He then fit the dimensions into three major clusters. Kulik and McKeachie (1975) reviewed 11 studies and identified four commonly found factors.

In sum, there appeared to be encouraging evidence regarding the processes of effective teaching. Descriptions of teaching by students appeared to fit into a reasonably finite set of categories. Faculty and students showed reasonable agreement as to the characteristics they considered important. When students rated faculty on these characteristics, groups of items formed into factors that reviewers were able to organize further. In light of such findings, it seemed reasonable to ask students to rate faculty to measure teaching processes. After all, students had the greatest exposure to faculty teaching and should be in a good position to judge.

Such thinking, however, depended first on showing that students were accurate, consistent, and unbiased judges. Second, it depended on showing that the teaching qualities students were asked to judge were always relevant and appropriate and took into account innovative teaching methods[1] It also depended on showing that different rating forms contained items which tapped the same teaching qualities. Finally, it depended on showing that the results of specific ratings could be effectively used.

The accuracy of student ratings. The validity of student ratings as process measures of effective teaching depends on showing that the ratings of students are accurate and reliable descriptions of preparation and delivery activities. The reliability of student ratings is not a contested issue: the stability of ratings over time and the consistency of ratings over students (especially in classes of ten or more) compares favorably with the best objective tests (Feldman, 1977; Marsh, 1987; Marsh and Dunkin, 1992).

The accuracy of student ratings of teaching process is a concern about criterion-related validity. Are students able to accurately judge whether (quantity) and how well (quality) instructors teach according to the dimensions specified on the rating form? In general,

[1] Feldman (1976, 1988, 1989a, 1989b, 1990) among others has explored the relationship between global ratings of teaching effectiveness and dimensional ratings as a way of showing the validity of dimensional ratings as indices of teaching processes. The value of such an approach depends on making a case for the link between specific teaching processes and student perceptions of the general quality of teaching received. Feldman (1988) puts the case this way:

> If it is assumed that each student's overall evaluation of an instructor is an additive combination of the student's evaluation of specific aspects of the teacher and his or her instruction, weighted by the student's estimation of the relative importance of these aspects to good teaching, then it would be expected that students' overall assessment of instructors would be more highly associated with instructor characteristics that students generally consider to be important to good teaching than with those they consider to be less important. (p. 314)

The assumption of a link between global ratings and specific ratings is, in our view, highly plausible but an assumption that can be challenged on both conceptual and empirical grounds. Is it not also plausible that students' impressions have either: a) a general component and specific components or b) only specific components? If either of these alternative views is plausible, it would be erroneous to invalidate ratings of teaching dimensions that do not correlate with global assessments. For example, the social psychology literature suggests several models of impression formation including the three dimensions of evaluative judgment (good-bad, weak-strong, and fast-slow) offered by Osgood, Suci, and Tannenbaum (1957) as well as the weighted averaging model of overall impressions (Anderson, 1968).

However, what is fundamentally important is not the structure of student impressions but the structure of what teachers actually do when they teach. Consequently, the plausibility of the assumption that students form general impressions may be reasonable for the *judgment* of teaching process that students utilize but the assumption becomes much less reasonable and much less plausible when one is utilizing student ratings to develop a theoretical *description* of the teaching process. Is teaching a series of discrete actions? Do these actions meld into a single collection of actions or several collections of actions? It remains uncertain which of these ways is best to describe teaching.

criterion-related validation studies require alternative measures of the teaching process in addition to student ratings. For example, to assess the criterion-related validity of ratings as process measures requires examining studies comparing faculty (peer) and chair ratings with student ratings, trained observers ratings with student ratings, instructor self-ratings with student ratings, etc. The data suggest that students are reasonably accurate judges of most teaching processes (Marsh, 1987; Marsh and Dunkin, 1992).

The criterion validation of student ratings as measures of teaching processes is not to be confused with the validation of student ratings as measures of teaching products. As Doyle noted:

> In instructional evaluation validity studies, ratings of instructor characteristics are compared with student learning. But student learning is not an alternative measure of, say, an instructor's effectiveness in engaging student attention. Alternative measures of engaging student attention might include observer's counts of students dozing or staring out the window, or student reports of boredom, or even galvanic skin response. (1981, p. 24)

The content validity of student ratings. The validity of student ratings as process measures of effective teaching also depends on showing that the items on the rating form have content validity and are a representative sample of items from the larger population of items. The requirement of content validity suggests that if a single form is used it is equally applicable in a variety of instructional contexts and not just the lecture format for which most rating forms were designed. These instructional contexts include different pedagogical methods (e.g., small and large class lecturing, tutoring and advising, studio classes, discussion and small group methods including cooperative learning, individualized and mastery learning, etc.), academic disciplines, student and setting characteristics, etc.

Abrami, d'Apollonia and Cohen (1990) argued that a student rating form should contain items equally relevant to each of the instructional situations for which it was designed. Consequently, items such as "Students were encouraged to participate in class discussion" and "Instructor was friendly towards individual students" would not be equally relevant in small and large classes, regardless of whether those items retained the same interrelationship with other items across instructional contexts. For example, imagine several items (e.g., friendliness, openness, encouraging, and warmth) which assess instructor rapport. It is quite easy to see how scores on these items would be interrelated regardless of instructional context. If you are not very friendly, you are probably not seen as especially open, encouraging, or warm. This may explain why Marsh and Hocevar (1984, 1990) report some evidence of the factorial validity of the SEEQ.

But it is equally easy to envision how instructor rapport with students might be more critical in a small class than a large one. And it is also possible that because different teaching behaviors are important in different contexts, instructor mean ratings might vary across contexts. That is, instructors may concentrate on the qualities important in that context and receive higher ratings on context-relevant teaching skills. Fernald (1990) found that items on a multidimensional rating form varied greatly with regard to student perceptions of item relevance to the course. Furthermore, the degree of item relevance was correlated with student ratings of instruction: the higher the item relevance score, the higher the student rating score.

In our hypothetical example, rapport mean ratings would vary significantly in small classes versus large classes even though the underlying relationship among rapport items remained the same. Unfortunately, mean scores, not interitem correlations, are used by promotion committees to make summative decisions about effective teaching. In this case, irrelevant items bias the case against the instructor of large classes.

Comprehensiveness and uniformity of student rating forms. Another type of evidence concerning the validity of rating forms comes from comparisons of items on different rating forms. Abrami, d'Apollonia and Cohen (1990) reasoned that if effective teaching was substantially invariant, then one would expect the same teaching qualities to emerge on each multidimensional rating form; there would not be substantial variability across forms in the factors of effective teaching which are assessed. Moreover, the relative type and proportion of items representing these factors would also not vary across forms.

To assess the comprehensiveness and uniformity of existing multidimensional rating forms, we used an early version of our coding scheme to sort the rating items found in 43 studies assessing the validity of student ratings to predict teacher-produced student learning. There were 154 study findings in the 43 studies (e.g., studies that report the findings for more than one course). There were 742 validity coefficients or correlations between scores on the rating forms and student learning. For example, a multidimensional rating form would yield several rating-achievement correlations. We first determined the number of times a category was found in the study findings. The comprehensiveness index represents the portion of times a teaching category is represented in the 154 findings. We then computed a uniformity index, which is a measure of the unidimensionality of reported validity coefficients across forms, for each instructional dimension. The uniformity index is the average proportion of items within a specific dimension, computed across 154 study findings. Thus, a high uniformity index indicates that the reported validity coefficients tend to represent a single dimension. The results of the uniformity and comprehensiveness analyses are presented in Table 2. The results suggest that both the items that appear on multidimensional student rating forms and the factors that these items represent vary across study findings.

Table 2: Uniformity and Comprehensiveness Analysis of Student Rating Forms (N = 154 study findings)

Dimension	N	CI[1]	UI[2]
Stimulation of interest	88	0.57	0.25
Enthusiasm	30	0.19	0.23
Knowledge of the subject	43	0.28	0.36
Intellectual expansiveness	35	0.23	0.11
Preparation and organization	89	0.58	0.33
Clarity and understandableness	112	0.73	0.30
Elocutionary skills	54	0.35	0.13
Class level and progress	76	0.49	0.20
Clarity of course objectives	68	0.44	0.25
Relevance and value of materials	46	0.30	0.38
Supplementary materials	26	0.17	0.36
Workload	84	0.55	0.45
Perceived outcome	75	0.49	0.47
Fairness of evaluation	69	0.45	0.39
Classroom management	79	0.51	0.25
Personality characteristics	54	0.35	0.25
Feedback	66	0.43	0.24
Encouragement of discussion	90	0.58	0.35
Intellectual challenge	35	0.23	0.24
Concern and respect for students	75	0.49	0.22
Availability and helpfulness	68	0.44	0.29
Overall course	92	0.60	0.51
Overall instructor	109	0.71	0.61
Miscellaneous	47	0.31	0.23

[1] CI =Comprehensiveness Index [2] UI =Uniformity Index
Adapted from Abrami, d'Apollonia and Cohen (1990).

The uniformity indices for teaching dimensions were as low as 0.11 (instructor expansiveness ratings); these dimensional indices contrast with global indices that were 0.51 (overall course rating) and 0.61 (overall instructor rating). Especially at the level of asking specific questions about instruction (i.e., low-inference questions) multidimensional student rating forms are composed of a diverse collection of items. Furthermore, even as the items are organized into factors, a considerable lack of uniformity remains.

Student ratings and innovative teaching methods. As never before, college instructors are using innovative teaching methods in place of, or in addition to, the traditional lecture method. One method that shows special promise for enhancing student achievement as well as developing communication and interpersonal skills, is cooperative learning (Abrami et al., 1995; Cooper et al., 1990: Johnson, Johnson, and Smith, 1991). Cooperative learning relies on students learning actively and purposefully together in small groups. Two key elements of cooperative learning are positive interdependence and individual accountability. Positive interdependence exists when students perceive that their success at learning has a positive influence on their teammates' successes and vice versa. Individual accountability exists when students perceive that they are responsible for their own learning and for the learning of their teammates. The instructor's role in cooperative learning is different than in whole class instruction. Because students spend a considerable amount of time attending to their classmates, much less class time is devoted to lecturing. Instead, the instructor usually gives only a brief overview of important ideas and then allows student teams to explore these ideas further.

The distinctiveness of cooperative learning compared with lecturing suggests that the specific instructional processes involved will be different. For example, in whole class instruction almost all of class time is devoted to the instructor talking and students listening. Clarity of explanation should be more important in classes designed for lecturing than in classes where the instructor presents for only a portion of the time.

In a cooperative classroom, the instructor's primary role is to insure that teams are viable and that teammates are effectively instructing one another. In particular, the instructor insures that group tasks are appropriate for learning and that students are operating together as a team with each member of the team holding a personal stake in the outcome. Furthermore, the instructor insures that each team has the necessary skills and abilities to learn. When necessary, the instructor may intervene to motivate students and to facilitate their learning. Thus, differences in instructional methods suggest that a student rating form consisting of one set of specific teaching dimensions will not have uniform content validity.

Factorial invariance? An underlying assumption of the multidimensional approach to the evaluation of instruction is that the characteristics of effective teaching are substantially invariant across situations (Marsh and Hocevar, 1984). In general, the qualities important to effective teaching are not expected to vary from course to course, from department to department, or from university to university. Marsh and Hocevar (1984, 1990) provide some evidence of the factorial invariance of one student rating form across different groups of students, academic disciplines, instructor levels, and course levels. That is, the factor structure of the rating form (i.e., the number and nature of the teaching dimensions found) and thus the relationships among perceived characteristics of teaching, was stable across contexts. However, differences in pedagogical methods were not explored for possible influences on factor structure.

A study by Smith and Cranton (1992) reached different conclusions about the influence of course characteristics. They found that student perceptions of the amount of improvement needed in the four dimensions of a student rating form differed significantly across levels of instruction and class size. They concluded that the relationships between course characteristics and student ratings are not general but specific to the instructional setting. They suggested several practical implications of their results. First, for instructional improvement, a faculty member should not assume that all items on a student rating form are of equal importance in planning changes. Second, faculty who want to determine criteria for the interpretation of their ratings by comparing themselves to others would likely be making a mistake. Third, personnel decisions using data from student ratings should not be based on a comparisons among faculty or across courses without considering the instructional setting.

Utility of student rating forms. Finally, one cannot expect untrained administrators or non-experts 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 (Franklin and Theall, 1989). One cannot expect administrators to have the expertise of faculty developers, nor are there precise and defensible procedures for synthesizing the information from factor scores. Experience suggests that administrators weigh factor scores equally or look for particularly strong or weak areas of teaching. What if these low scores occurred because the dimensions were low in relevancy? Cashin and Downey (1992) studied the usefulness of global items in predicting weighted composite ratings with a sample of 17,183 classes from 105 institutions. Their results were that global items accounted for a substantial amount of the variance (more than 50%). They concluded: "The results of this study have supported that single, global items—as suggested by Abrami (1985)—can account for a great deal of the variance resulting from a weighted composite of many multidimensional student rating items" (Cashin and Downey, 1992, p. 569). They recommended that short student rating forms should be used for summative evaluations and longer forms should be reserved for teaching improvement.

Ratings and the processes of instruction: Where do we go from here? The interests of many researchers and practitioners alike appears to have focused on finding a rating form capable of identifying the major qualities or traits essential to the process of effective teaching. Analytical strategies such as factor analysis concentrate on identifying what is common to teaching and generally disregard what is unique.

The alternative view we argue for here suggests that the search for a collection of the invariant dimensions of effective instruction may underemphasize the importance of the local context. We are reminded, in particular, of the endless discussions among faculty over the merits of including particular items on student rating forms. Comments such as "What does_____have to do with good teaching?" are reflections of the possible problems associated with employing a single definition of instruction when many are needed. Consequently, research and practice may need to be more sensitive to situational influences and make greater allowances for multiple approaches to the definition and evaluation of effectiveness.

Nevertheless, there are both theoretical and practical reasons to continue to examine, describe, and classify instructional processes. We decided, therefore, to explore further the research on the dimensionality of the processes of effective teaching by quantitatively integrating the results of many studies using a collection of different student rating forms. We believe that a systematic effort to integrate this corpus of research may better answer questions about teaching. Is there a core set of teaching qualities that emerge from every one of the studies? Do these qualities form into the same factors? How much does context matter? By integrating the existing research, we hoped to be better able to separate common dimensions of teaching from unique qualities that may only be appropriate for particular instructional context. We describe our findings in a later section. Before doing so, we consider research linking the processes and the products of effective teaching.

EMPIRICALLY DETERMINING THE LINKS BETWEEN THE PROCESSES AND PRODUCTS OF EFFECTIVE TEACHING

According to the process-product view of effective instruction, a valid student rating must assess accurately, if not directly, instructor impacts on both processes and products. That is, we wish to know not only the extent student ratings reflect what instructors do when they teach but also the extent to which students learn course content, are motivated, and develop critical skills as a result. Consequently, the principal consideration for a research design is that it allows one to assess the degree to which student ratings reflect what teachers do (process) and the impact teachers have on students (product). In particular, the design must control for plausible rival explanations to the causal effects of instructors.

Generally, these plausible rival explanations center around the effects of "biasing" characteristics, mainly student characteristics (e.g., ability), but also course and setting effects (e.g., size), and extraneous instructor characteristics (e.g., grading standards). Thus, our first consideration is that the design controls for plausible threats to internal validity (Campbell and Stanley, 1963).

Our second consideration is that the design allows us to generalize the results across students, instructors, courses and other setting characteristics, various rating instruments and importantly, different products of effective instruction. For example, we wish to conclude that ratings predict teacher impacts in a variety of courses and for a variety of instructor effectiveness measures. Thus, our second consideration is that the design controls for plausible threats to external validity (Campbell and Stanley, 1963). The strongest design will control for plausible threats to both internal and external validity.

In this section three research designs—the laboratory design, the multisection validation design and the multitrait-mutimethod design (MTMM)—are critically reviewed. In the typical laboratory design, students are randomly assigned to instructional treatment conditions that attempt to simulate certain classroom features. After a brief exposure to the treatment (often as short as 20 minutes), students are asked to complete ratings and other measures. In the multisection validation design, researchers correlate mean student ratings and mean student achievement on a common examination from multiple sections of a college course. A large positive correlation is taken as evidence of rating validity, establishing a link between what instructors do when they teach and their impact on students. In the MTMM design, student ratings factors and several criterion measures (e.g., instructor self-ratings) are collected across a wide range of courses, and the convergent and discriminant validity of ratings are assessed.

Laboratory Designs

To explore simply and conclusively the causal relationships between particular instructional processes and particular products requires experiments that manipulate what teachers do and that measure how students change as a result (see Murray, 1991, for a review). Laboratory designs are the strongest designs for controlling threats to internal validity because they manipulate instructional conditions and control for the effects of students through random assignment. However, they are the weakest designs for controlling for threats to external validity.

The laboratory studies on instructor expressiveness and lecture content (educational seduction or the Dr. Fox effect; Abrami, Leventhal, and Perry, 1982) examined the effects of two instructional delivery processes—expressiveness and content—on two instructional products—student satisfaction and low-level student learning. But Abrami, Leventhal, and Perry (1982) argued that these laboratory studies suffered shortcomings in both the comprehensiveness of the process variables studied and the representativeness of the values of the process variables manipulated. The laboratory studies lacked comprehensiveness because they failed to represent the many instructor characteristics that may affect ratings and learning. The laboratory manipulations of instructor characteristics lacked representativeness because they failed to represent actual differences among instructors in the field. The lack of both comprehensiveness and representativeness means that laboratory studies cannot be used to estimate the *extent* to which ratings predict student learning. For example, the laboratory findings that instructor expressiveness affects ratings substantially ($r = .70$) and achievement slightly ($r = .12$) suggests only that the correlation between ratings and achievement falls somewhere in the range of $+.84$ to $-.56$. Instead, laboratory studies are best used to explain *why* ratings and achievement are related by identifying the instructional processes which *causally* affect instructional products.

MULTISECTION VALIDATION DESIGN

To date, more than 40 studies have appeared using the multisection validation design. The design has several features that make it high in internal validity. Using class section means rather than students (or students pooled across classes) as the units of analysis emphasizes instructor effects on ratings and achievement. Furthermore, in many of these studies, section differences in student characteristics were controlled experimentally, via random assignment, or statistically, using ability pretests. Similarly, section differences in setting effects were often minimized with the use of a common syllabus, common textbook, similar section sizes, and so on. Finally, the effect of instructor grading standards was reduced by the use of a common examination for all sections. Thus, the design minimizes the extent to which the correlation between student ratings and achievement can be explained by factors other than instructor influences. However, unlike laboratory studies, instructional variables are not manipulated but only measured by the student rating instrument.

One of the strongest features of the design is that the validity criterion, mean section examination performance, is relatively high in external validity. Examination scores are both a direct and important measure of one of the products of effective instruction, designed to assess what students have learned of the course material (and to assign grades). Consequently, we believe that multisection validation designs are especially useful in determining the extent to which ratings of particular instructional processes are valid indices of important instructional products, particularly student learning of course content.

Substantive criticisms of multisection validation designs. Feldman (1989a, 1990) expressed a different view of the value of multisection validity studies:

> Although the data for the present analysis comes from what are called "multisection validity studies," the analysis herein was not an attempt to validate specific ratings of instructors. While it makes sense to seek information about the validity of overall or global ratings of instructors by correlating these ratings with student achievement, it makes less sense to do so for specific ratings because student achievement is not necessarily a direct or meaningful validity criterion for each of the instructional dimensions…

> The present analysis accepted the specific rating items, scales, and factors of the studies under review as valid indicators of instructional characteristics. It sought to find out which of them are most highly associated with student achievement under the presumption that the higher the correlation the more facilitative is the instructional characteristic of student achievement. (1989, pp. 624–625)

We do not share completely Feldman's interpretation of the value of multisection validity studies. We agree that understanding the relationship between global ratings and student achievement is extremely important and can be used in judging the validity of global ratings. However, we believe that understanding the relationship between specific ratings and student achievement is also important and can be used to judge the validity of specific ratings since it sheds light on the link between what instructors do when they teach and their impact on students. According to the process-product view, ratings dimensions *are* validated to the extent they reflect instructor-produced student learning.

Methodological criticisms of the multisection design. Abrami (Abrami, Cohen, and d'Apollonia, 1988; Abrami, d'Apollonia and Cohen, 1990) and Marsh (1987; Marsh and Dunkin, 1992) disagree over the strengths of the multisection design. Marsh gives several reasons why the design of multisection validity studies is "inherently weak" and notes that "there are many methodological complications in its actual application" (1987, p. 289). First, the sample size of course sections in any study is almost always quite small, adversely affecting sampling error. Second, variance in achievement scores is mostly attributable to student variables (e.g., ability) and researchers are generally unable to find appreciable effects due to teachers, especially in multisection designs where many of the setting effects are held constant. In addition, the reliability of section average differences is unstudied but may be small and unreliable, attenuating the size of the ratings-achievement correlation. Third, the comparison of findings across different multisection validity studies is problematic since most use different operationalizations both of student ratings and achievement. Fourth, other criteria of teaching effectiveness besides objectively scored tests, and more generally student learning, need to be considered. Fifth, pretest scores on student ability should be used to statistically equate course sections even when students are randomly assigned to the sections, since randomization is not a guarantee of section equivalence. Furthermore, the multisection design does not constitute an experimental design in which students are randomly assigned to treatment groups that are varied systematically in terms of experimentally manipulated variables, and so the advantages of random assignment are not so clear. Finally, the grading satisfaction hypothesis may explain the ratings-achievement correlation. According to the grading satisfaction hypothesis students reward teachers who assign high grades by rating instructors highly regardless of how much students actually learned.

Response to criticisms of the multisection design. We agree with Marsh on several points. First, integrating the findings from the collection of multisection courses helps overcome sample size problems in analyzing single studies. The research we report here and elsewhere is an attempt at such integration. Second, the statistical control of student characteristics in combination with randomization can be superior to randomization alone. However, failing this and faced with a choice of design strategies, we prefer the use of experimental control of nuisance variables over statistical control for two reasons: a) the as-yet unstudied effect of poor randomization on the validity coefficient must certainly be less than when students self-select course sections; and b) statistical control requires that these nuisance variables are known and uncorrelated with instructor effects, while random assignment does not.

Third, we agree that the products of effective instruction are multidimensional. But a call for the inclusion of measures other than student learning is not, by itself, an identification of a methodological weakness in multisection designs. It does identify a limitation of existent studies and suggests a direction for future research. Furthermore, the learning measures studied in multisection investigations do represent multiple operationalizations of student learning since test item content varies from study-to-study. Finally, instructor self-ratings, colleague or peer ratings, and the ratings of trained observers could be incorporated into studies employing the multisection design.

We disagree with Marsh on several points. First, the restriction of range problem in the achievement criterion does not hold unless it can be shown that the sample of instructors studied is unrepresentative and the criterion measure lacks sensitivity to instructor effects. Otherwise the experimental control of extraneous influences which affect the criterion is desirable, not undesirable. In both laboratory and field investigations, Abrami (Abrami, Perry, and Leventhal, 1982; Abrami and Mizener, 1985) found that student ratings were more sensitive than student achievement to differences in instruction. Instructors may have genuinely small effects on what students learn. In addition, the use of locally developed or teacher-made tests in some of the validation studies is a double-edged sword. On the one hand, teacher-made tests are likely to be less psychometrically sound but, on the other hand, are often more sensitive to instructor effects than standardized tests.

Second, mono-operationalizations of measures (i.e., using the same instruments throughout) reduce, but do not eliminate, the interpretive problems involved in making inferences across multisection validity studies. Important uncontrolled differences in student, instructor, course, and setting characteristics may also be responsible for study-to-study differences therefore lowering the internal validity of cross-study comparisons. However, cross-study comparisons can be useful for judging the external validity of findings where it seems reasonable to explore whether different student ratings instruments are correlated with different student learning measures.

Third, the unsystematic nature of the treatment (i.e., differences in instruction) in multisection designs does not detract from the value of random assignment of students. Random assignment helps insure that the relationship between ratings and achievement was produced by differences in instruction rather than differences in students. This insurance of internal validity can be the starting point for further explorations of the treatment. For example, Sullivan and Skanes (1974) used the multisection design to explore the influence of instructor experience on the ratings-achievement relationship.

Finally, the grading satisfaction hypothesis may be one mechanism by which students rate faculty, but it is not an alternative explanation of the validity of ratings when section differences in students are controlled and instructor grading practices, including timing, are uniform across classes. The alleged effect of grading satisfaction will operate consistently, if at all, in each section of a multisection course unless instructors *first* produce differences in student learning. Under these conditions, grading satisfaction cannot explain mean section differences in either student ratings or student achievement. However, the problem is especially pronounced when one is studying multiple classes outside the multisection paradigm where there is more variability in instructor grading practices.[2]

[2] Marsh and Dunkin (1992) suggest several fallacies with our reasoning: First, the implicit assumption that all section differences are instructor-produced is completely unrealistic. Even random differences in section mean examination performances will produce inflated validity

MULTITRAIT-MULTIMETHOD DESIGNS

Marsh (1987) and others (Howard, Conway, and Maxwell, 1985) have argued in favor of a MTMM approach to the validation of ratings. To be superior to the multisection design, the MTMM design requires greater control of threats to internal validity, external validity, or both. Specifically, the design must reasonably show that threats to internal validity are controlled in order to attribute class mean differences in ratings and the criterion measures to instructors, and not to extraneous characteristics such as students, the course, and setting variables.

coefficients due to grading satisfaction effects. Second, there is no way to unconfound the influence of grades and satisfaction with grades. Third, the problem *is* pronounced outside the multisection paradigm and, therefore, the paradigm is unrepresentative. Finally, the reliability of section-average differences in achievement is a critical problem when the section-average scores are similar to one another and within-section differences.

Our response follows: First, if there were unwanted systematic differences in section means they would be much smaller than validation designs without controls for student differences. (The point of our argument has always been that the multisection design is relatively one of the strongest designs, not that it is a perfect design.) It is also unclear what effect *random* differences in examination performance will have on the ratings-achievement relationship. In general, unsystematic differences tend to attenuate the size of a correlation, whereas Marsh and Dunkin (1992) claim the opposite will occur due to grading satisfaction. It is also the case that inferential statistics were conceived on the notion of random fluctuation both between and within groups. This random fluctuation or sampling error does not need to be zero for valid statistical tests to be performed, although reductions in error variability increase the power or sensitivity of the tests. Random or unsystematic fluctuation, a tolerable problem, is not to be confused with systematic bias or contaminants which are alternative explanations of teacher effects, a more serious problem.

Second, we agree that the grading satisfaction effect cannot be disentangled from the effect of grading per se in existing multisection studies although it could be incorporated into the design of future multisection studies. Our claim is that the temporal sequence of influence (i.e., instructor produced learning affects grades which affects satisfaction which may affect ratings) coupled with the use of uniform grading standards removes grading satisfaction as a source of bias. Marsh's claim is the influence of learning and grade satisfaction on ratings are dissimilar. For example, small differences in learning produce large differences in grading satisfaction which, in turn, have a meaningful impact on student ratings. Thus, if this were the case it would be seen in individual validation studies incorporating a grading standards variable or by comparing multisection validation studies where grading standards were not uniform with validation studies where grading standards were uniform.

Finally, we believe that Marsh's concern for the reliability of section mean differences in achievement should be extended both to all student ratings and criterion measures and to all designs, not only multisection validation designs, using the correct unit of analysis for exploring instructor influences which is the class mean or section average. For example, in 1990 we wrote: "if we adjusted a validity coefficient of .43 (which is the average value reported by P.A. Cohen, 1981, for overall instructor ratings) for the reliability of ratings (estimated to be .70), the corrected coefficient would be .51. If we then adjusted the validity coefficient further for an equal degree of error in the criterion measure, the corrected coefficient would be .61" (Abrami et al., 1990, p. 227).

This can be partly achieved if the criterion measures of effective instruction—possibly instructor self-ratings, alumni or former student ratings, peer ratings, and ratings of trained observers—are less sensitive to extraneous influences than course examinations. If so, one may compute the validity correlation between student ratings and scores on the criterion measure(s) for a host of courses, not just multisection ones, and may thereby greatly enhance external validity. But without evidence to the contrary, designs which do not control statistically or experimentally for extraneous influences on the criterion do not represent good alternatives to the multisection validation design in concluding that differences in the criterion measure were caused by instructors. Furthermore, these designs do not control for extraneous influences on student ratings. Thus, even if it could be shown that the criterion was unaffected, the validity coefficient might be affected.

To show advantages in external validity, one must also show that the alternatives to student learning such as instructor self-ratings, former student ratings, and peer ratings represent adequate product measures of instruction. Yet whether these measures (and student ratings) represent adequate criteria of effective instruction has been seriously questioned (Gaski, 1987). Maxwell and Howard (1987) acknowledge these criticisms as well-taken (see also Feldman, 1989b). In our view, such measures help establish the validity of ratings as measures of teaching processes but not as measures of the products of instruction.

Thus, we conclude the MTMM validation designs provide weaker evidence for the validity of student ratings as measures of instructional effectiveness than multisection validation designs. The MTMM designs are generally weaker in internal validity and employ criterion measures which are either less defensible as or less important measures of good teaching than student learning.

The Choice of Designs

In choosing among research designs, one must consider whether threats to internal and external validity are addressed. The multisection validation design has advantages over MTMM designs in determining whether ratings reflect instructional processes and products. The multisection design is generally higher in internal validity and typically incorporates an important product measure of effective instruction, student learning, contributing to its external validity. The multisection design is also superior to laboratory studies when the validity question addresses the practical concern of the degree to which ratings predict teacher-produced outcomes in typical classroom settings. For these reasons, multisection validation studies are singularly important to concerns about validity and deserve special attention.

MULTISECTION VALIDITY STUDIES: WHAT HAVE THEY TOLD US SO FAR?

What can one conclude about the validity of ratings from multisection validation studies? Do global ratings predict student learning? Are there particular instructional processes, as reflected in student ratings, which are related to student learning or other outcomes? Are the findings from the collection of studies uniform? If not, are there substantive or methodological features that explain variability in study findings? Do reviewers agree on what the findings mean? Is there more to learn: Are there inadequacies in either the literature or reviews of the literature?

THE RELATIONSHIP BETWEEN RATINGS AND STUDENT LEARNING

Abrami, Cohen, and d'Apollonia (1988) compared six published, quantitative reviews of the findings from multisection designs (Abrami, 1984; Cohen, 1981, 1982, 1983; Dowell and Neal, 1982; McCallum, 1984) to identify their agreements and disagreements. Unfortunately, the reviews differed in several important ways including: a) the specification of the criteria used to include studies; b) comprehensiveness or the extent to which each review included studies meeting inclusion criteria (where the proportion of studies included per review ranged from .13 to .88); c) the presence and completeness of study feature coding used to explain study-to-study variability; d) the extraction and calculation of individual study outcomes (where there was only 47% agreement among the reviews); and e) procedures for data analysis, especially variability in study outcomes. These difference help explain why the conclusions reached by the reviewers were markedly different:

> The present meta-analysis provides strong support for the validity of student ratings as measures of teaching effectiveness. Teachers whose students do well on achievement measures receive higher instructional ratings than teachers whose students do poorly. This study demonstrates that the relationship between ratings and achievement is slightly stronger and more consistent than was previously thought. (Cohen, 1981, pp. 300–301)

> The literature can be seen as yielding unimpressive estimates of the validity of student ratings. The literature does not support claims that the validity of ratings is a consistent quantity across situations. Rather the evidence suggests that the validity of student ratings is modest at best and quite variable. (Dowell and Neal, 1982, p. 59)

There have been further attempts to summarize the findings from the multisection validity studies and analyze variability in study findings (Abrami and d'Apollonia, 1987, 1988; Abrami, d'Apollonia and Cohen, 1990; Cohen, 1986, 1987; d'Apollonia and Abrami, 1987, 1988; Feldman, 1989a, 1990). The average validity coefficients found by the reviewers using two different coding schemes for categorizing the results from different rating forms are presented in Tables 3 and 4.

Collectively, the results of the reviews suggest that some specific rating dimensions, as well as student global ratings, are moderately correlated with student learning in multisection college courses. On average, there exists a reasonable, but far from perfect, relationship between some student ratings and learning. To a moderate extent, student ratings are able to identify those instructors whose students learn best. Furthermore, regardless of the coding scheme used, the average of global ratings of instructional effectiveness explains a greater percentage of variance in student learning than the average of specific ratings. It also appears that not all specific ratings are related to achievement; for example, ratings of course difficulty generally do not predict student achievement at all. Consequently, we recommend

Table 3: Mean Validity Coefficients in the Multisection Validity Studies: Cohen Dimensions (Cohen, 1987)

Type	Dimension	N[1]	VC[2]	Mean	SE[3]	Range
Global	Overall Instructor	59	0.44	0.45	0.012	[0.44, 0.48]
	Overall Course	21	0.48			
Specific	Skill	44	0.41	0.34	0.053	[0.00, 0.55]
	Rapport	35	0.30			
	Structure	29	0.55			
	Difficulty	25	0.00			
	Interaction	20	0.45			
	Feedback	7	0.29			
	Evaluation	25	0.23			
	Learning Progress	17	0.46			
	Interest/Motivation	12	0.26			

Table 4: Mean Validity Coefficients in the Multisection Validity Studies: Feldman Dimensions (d'Apollonia, and Abrami, 1988)

Type	Dimension	N[1]	VC[2]	Mean	SE[3]	Range
Global	Overall instructor	44	0.30	0.32	0.019	[0.30, 0.36]
	Overall course	18	0.36			
	Stimulates interest	34	0.37			
	Enthusiasm	11	0.25			
	Knowledge	12	0.21			
	Expansiveness	4	0.03			
	Preparation	33	0.43			
	Clarity/understandable	46	0.42			
	Elocutionary skills	8	0.26			
	Concern for progress	19	0.30			
	Clarity of objectives	25	0.33			
	Course materials	19	0.29			
	Supplemental materials	12	0.17			
Specific	Perceived outcome	32	0.39	0.20	0.015	[0.03, 0.45]
	Instructor's fairness	29	0.31			
	Personality	4	0.45			
	Feedback	15	0.11			
	Openness	27	0.29			
	Intellectual challenge	11	0.34			
	Concern for students	26	0.24			
	Availability	22	0.30			
	Course difficulty/workload	29	0.03			
	Classroom management	13	0.13			
	General	16	0.31			

using the results of specific rating dimensions to judge which teachers best promote student learning with caution especially when making promotion and tenure decisions. The same caution is not necessary when using global ratings of instruction.

Finally, the nature and number of the specific rating dimensions used in the two schemes appears different. In the Cohen (1987) coding scheme, the findings are arranged according to two global dimensions and nine specific dimensions. This coding scheme is not without limitations. For example, it relies on the factor analytic findings from a single instrument (Isaacson et al., 1964) which may not allow the results from all instruments to be properly represented. This may have resulted in validity coefficients being either forced into categories, creating heterogeneous categories, or dropped from the meta-analysis.

d'Apollonia and Abrami (1988) used Feldman's scheme to report the average validity coefficients for 22 specific rating dimensions, more than twice the number reported by Cohen. This coding scheme has been used and refined repeatedly by Feldman (1976, 1983, 1984, 1989a) using a conceptual approach to comprehensively represent the items from many forms without preference toward any one instrument or its factor structure. Nevertheless, questions remain about which way to organize items and whether any coding scheme can be empirically validated.

d'Apollonia and Abrami (1988) and Abrami et al. (1990) also observed that the multisection studies contained a large number of validity coefficients that were not all represented when other reviewers reported mean coefficients. In 43 multisection validation studies we found a total of 742 ratings-achievement correlations reported. Yet only a small fraction of these correlations were included in other reviews.

Looking Further at the Multisection Validity Studies

Research integrations are seldom necessary when the findings in an area are uniform. Reviews become necessary especially when the results of research on a topic appear heterogeneous. The research findings from the multisection validity studies seem to vary widely. The range of reported validity coefficients is −0.75 to +0.92. There is one study finding of a strong negative relationship between ratings and achievement—the highest rated instructors had the lowest performing students. There is also one study finding showing the opposite, a near perfect positive relationship between ratings and achievement. In a quantitative review, the reviewer searches for ways to explain the variability in study findings. The range of findings in the multisection validity studies is a reason to explore the findings further to explain these inconsistencies.

Cohen (1981) was the first quantitative reviewer to attempt a systematic exploration of the variability in study findings. He explored the relationship between 20 study and methodological features of the primary research and the validity coefficients extracted from this research. Three of the features together accounted for approximately thirty percent of the variance in the validity coefficients for Overall Instructor ratings: control for bias in evaluating achievement (i.e., Was the test graded by the instructor?); time at which ratings were administered (i.e., Were the ratings collected before final grades?); and instructor experience (i.e., Were the instructors graduate students?).

Abrami et al. (1990) examined these findings further. First, they showed that individual study features did not explain a significant amount of variability in the validity coefficients because of low statistical power. The analyses often lacked the sensitivity necessary to identify characteristics that explain a medium size effect on the relationship between ratings and achievement. More of the study features might prove to be useful predictors in future if either additional primary studies are conducted or more powerful statistical procedures of research integration were employed. Until then, one can neither accept nor reject claims that these other explanatory factors are trivial.

Second, the 20 study features employed by Cohen (1981) to explain variability in validity outcomes did not generalize across global and specific aspects of teaching. Since rating factors are regarded by some as distinct and uncorrelated (e.g., Marsh, 1987) there is little reason to suspect that these factors will be uniformly affected by biasing characteristics. Characteristics that predicted the relationship between student perceptions of teaching and instructor impacts on learning varied with the aspect of teaching being investigated. Unfortunately, the precise nature of this pattern of effects could not be elaborated. (For one, the sample sizes were too small to confidently make fine distinctions.) However, the findings were sufficiently clear to urge users of multidimensional rating forms away from the common practice of universally controlling for "biasing" characteristics (e.g., course level) and further complicate the use of specific ratings in summative decisions about teaching.

Third, Abrami et al. (1990) employed nomological coding to identify the investigated, accounted for, and mentioned characteristics in the forty-three validity studies. They uncovered 75 study features that could be used to explain variability in the findings of the multisection validity studies. Since this was a substantial increase in the explanatory features used previously (almost four times the number of factors explored by Cohen, 1981), Abrami et al. (1990) concluded that prior reviews did not comprehensively identify potential predictive characteristics.

WHERE DO WE GO FROM HERE?

Reviews of the multisection validity studies on the relationship between ratings and achievement suggest there is much yet to be learned of the relationship between what instructors do when they teach and how this affects student learning and other products of instruction. Abrami et al. (1990) recommended that another quantitative review should be undertaken with alternative systems for coding the rating dimensions, the use of the 75 study features they identified, and more powerful analysis strategies (e.g., tests of homogeneity, Hedges and Olkin, 1985) using the 742 validity coefficients extracted from the literature. As a major step in this process, we first embarked on research to identify the common dimensions of teaching as represented in the rating forms used in the multisection validity studies. Before presenting the results of the integration of many forms, we discuss some of the complications with the factor analysis of a single form.

FACTOR ANALYSIS AND THE DIMENSIONS OF EFFECTIVE INSTRUCTION

Recent research by Marsh (1991) attempted to address questions about the dimensionality of student ratings through the application of confirmatory factor analysis (CFA). Using data from a single rating form—the SEEQ—Marsh (1991) evaluated four higher-order factor models. The results provided support for the nine first-order factors the SEEQ is designed to measure and, of the higher-order models, particularly for the four factor approach. Marsh concluded:

> Considerable information is lost when the student ratings are summarized with a single score or even a small number of scores. The challenge for future research—particularly in terms of personnel decisions—is how to most appropriately use the information that is available in student ratings rather than throw it away. (Marsh, 1991, p. 13).

The validity of this conclusion rests in large part on the adequacy of the SEEQ to represent the qualities of effective teaching. As is evident from Abrami, d'Apollonia and Cohen (1990), different student rating forms assess different dimensions of effective instruction. This point is also demonstrated by Marsh (1991, see Table 1, p. 15) where the dimensional categories of the Endeavor rating form (Frey, 1978) were compared with the SEEQ and shown to be different. It is also not surprising that Marsh's (1991) confirmatory analyses generally conform to prior analyses with the same instrument—they amount to grand tests of instrument reliability. What is surprising is that the nine factor a priori model… "is not fully adequate" (Marsh, 1991, p. 9).

In addition, note that Marsh (1991) was able to describe not one but four higher order models from prior research and reviews. But even these models do not adequately describe the diversity and complexity of findings regarding the dimensionality of instruction. In his reviews of student rating forms, Feldman (1976, 1988) noted that rating form items are often intercorrelated despite their apparent conceptual independence. For example, the instructor's stimulation of interest, clarity and comprehensibility, course preparation, and organization and enthusiasm are frequently highly correlated. Kulik and McKeachie's (1975) review of student ratings suggested a general Skill factor on which many items (including global items) load highly. Global items are included in the specific factors from the SEEQ: *Overall course rating* is included in the *Learning/Value* factor and *Overall instructor rating* is included in the *Instructor Enthusiasm* factor. These interrelationships among items and between global items and specific factors create interpretive difficulties when one argues for the dimensionality of instruction. Should one interpret this covariance to mean that specific dimensional ratings predict global ratings or that students' responses to specific items are influenced by their overall assessments? Indeed, Feldman's reviews have often been predicated on the assumption that specific dimensional ratings should predict student global ratings.

LIMITATIONS OF FACTOR ANALYSIS

Implicit in Marsh's conclusion is that CFA can "disconfirm" the theory that a few general or global dimensions capture the structure of student ratings. An alternative conclusion is that it is not the theory which needs revision but the instruments and methods used in testing it.

The use of factor analysis alone to determine the structure of a phenomenon is inconclusive since different analysis methods are based on different assumptions. Each analysis is, therefore, designed to "discover" the structure favored by the assumptions. For example, principal components extraction, the most frequently used extraction method, was pioneered by Spearman to extract mutually independent components such that the first or principal component resolves the maximum amount of variance with subsequent factors explaining progressively less variance. Thus, factor analysis without rotation is designed to resolve one general or global component explaining most of the variance and a few less important, subsidiary components. Thurstone objected to this hierarchical interpretation of the components and developed rotation to redistribute the variance explained by the general factor over the subsidiary factors. This increased the variance the subsidiary factors explained.

The two solutions resolve exactly the same amount of total variance, therefore, which one "best" describes reality cannot be determined empirically. Moreover, both solutions are affected by the choice of items in the instrument(s) in question. The selection of unique items that are highly positively correlated favors a principal components solution while the selection of clusters of similar items favors a rotated solution elucidating a number of equally important factors. Clearly, the use of any one student rating form in a CFA is not an adequate test for the presence of a particular higher order factor structure. An adequate test requires the use of a diversity of student ratings.

Secondary Analyses of SEEQ Data

There are a number of decisions made during a factor analysis that affect the final results and their interpretation. Some of these are: a) the items included in the correlation matrix, b) the number of factors extracted, c) whether the axes are rotated, d) if rotated, whether rotated orthogonally or obliquely, and e) if oblique rotation is selected, the degree of obliqueness. In order to investigate the possibility that Marsh's conclusions reflected his methodological choices rather than the multidimensionality of instruction (as determined by the SEEQ), Abrami and d'Apollonia (1991) conducted a secondary analysis of the SEEQ data from Marsh and Hocevar (1984) and replicated in Marsh (1991). Marsh and Hocevar (1984) obtained a nine factor solution using oblique rotation with delta set at approximately −2.0.

Abrami and d'Apollonia (1991) reconstructed the reproduced correlation matrix by premultiplying the factor pattern correlation matrix by the oblique factor matrix and postmultiplying it by the transpose of the oblique factor pattern matrix (Tabachnick and Fidell, 1983). Abrami and d'Apollonia (1991) estimated the observed correlation matrix by replacing the diagonal elements of the reproduce correlation matrix with 1's. Since the communalities were very high, they were able to replicate the results within rounding error. They then factor analyzed the correlation matrix of the 35 items using SPSS (SPSS Inc., 1990).

The results of the Abrami and d'Apollonia re-analysis (see Table 5) were: a) Thirty-one of the items load highly on the principal component (> .63) with Overall Instructor and Overall Course being the most highly loading items (both .94), indicating that the first component was a general or global factor. This global factor explained almost 60% of the total variance in ratings. The four remaining items, concerning course difficulty and workload, loaded heavily on a second component

Table 5: Nonrotated Factor Pattern Matrix with Six Components Extracted via Principal Components Analysis[1]

SEEQ Item	Factor loadings on first six components					
	I	II	III	IV	V	VI
Course challenging	.893	.392	.017	−.109	.037	−.056
Learned something valuable	.873	.262	.021	−.148	.133	−.080
Increased subject interest	.868	.122	−.013	−.226	.153	−.031
Understood subject matter	.760	−.182	−.094	−.167	.201	−.142
Overall course rating	.940	.176	−.032	−.085	.017	−.120
Enthusiastic about teaching	.886	.017	.039	−.082	−.308	−.073
Dynamic and energetic	.875	.104	.045	−.133	−.324	−.156
Enhanced presentation with humor	.787	.009	.039	−.158	−.315	−.152
Teaching style held interest	.884	.075	.012	−.147	−.249	−.208
Overall instructor rating	.941	.058	−.037	−.024	−.181	−.105
Explanations clear	.868	−.057	−.180	−.076	−.072	−.121
Materials prepared and clear	.857	.065	−.300	.052	−.064	−.060
Objectives stated and pursued	.855	.130	−.255	.125	.056	−.094
Lectures facilitated note taking	.649	.153	−.485	.118	−.158	.043
Encouraged class discussions	.741	−.390	.389	−.233	.104	−.068
Students shared ideas/knowledge	.688	−.479	.394	−.226	.136	−.046
Encouraged questions and answers	.852	−.303	.237	−.108	.047	−.070
Encouraged expression of ideas	.780	−.414	.349	−.131	.084	−.016
Friendly towards students	.769	−.370	.256	.258	−.097	.117
Welcomed seeking help or advice	.756	−.310	.246	.393	−.122	.184
Interested in individual students	.817	−.277	.261	.303	−.089	.117
Accessible to individual students	.678	−.180	.138	.471	−.108	.298
Contrasted implications	.803	.010	−.205	−.156	−.016	.418
Gave background of ideas/concepts	.819	−.024	−.237	−.201	.033	.398
Gave different points of view	.818	−.107	−.207	−.158	.060	.375
Discussed current developments	.743	.035	−.142	−.270	.030	.340
Examination feedback valuable	.776	−.061	−.163	.336	.064	−.182
Examination methods fair	.808	−.149	−.163	.328	.069	−.146
Exams emphasized course content	.794	−.064	−.242	.289	.071	−.187
Readings/texts valuable	.639	.168	−.045	.131	.563	−.018
Added to course understanding	.754	.175	−.007	.122	.476	−.098
Course difficulty	.333	.840	.181	.075	−.068	.046
Course workload	.336	.793	.351	.037	.034	.065
Course pacing	.238	.794	.182	.092	−.100	.037
Hours/week outside class	.305	.726	.384	.068	.065	.115
Factor eigenvalues	20.67	3.95	1.75	1.42	1.18	1.05
% variance explained	59.1	11.3	5.0	4.0	3.4	1.9

Source: Abrami and d'Apollonia (1991), p. 414.

which explained an additional 11% of the variance. Interestingly, Cohen (1981) found that course difficulty items were poor in construct validity, predicting student learning near zero. The remaining four components explained only 5%, 4%, 3%, and 2%, respectively, and did not contain any items that did not load heavily on one of the first two factors.

In response, Marsh (1991) claimed that the most serious problem with the critiques of Abrami and d'Apollonia (1991) and Abrami (1988, 1989a, 1989b) was the failure to operationalize criteria for unidimensionality or multidimensionality. Further, Marsh claimed that..." the most defensible approach to evaluating unidimensionality is to test the existence of one latent trait underlying the data" (Marsh, 1991, p. 417). Consequently, one purpose of our analysis of the collection of rating forms was to explore the underlying nature of student perceptions of instruction across many rating forms as the first step towards testing the existence of one global trait.

In the past, Abrami (1985, 1988, 1989a, 1989b) and Abrami, d'Apollonia and Cohen (1990, 1991) have been critical of the methodological and substantive difficulties with factor-analytic research on student ratings. These problems have not led us to deny that teaching is multidimensional—it clearly is—but to suggest that research to date does not justify the use of factor scores from a single instrument in making summative decisions about teaching effectiveness. By determining whether there exists a "common" core among the collection of rating forms used in multisection validation studies we believe we will take a step toward overcoming some of the limitations described above.

THE DIMENSIONALITY OF INSTRUCTION: IS THERE A "COMMON" CORE?

In this section we explore the unidimensionality-multidimensionality issue further by applying the techniques of quantitative synthesis to a collection of student rating forms. In this way, we will be able to explore the dimensionality of ratings with a higher degree of generalizability than ever attempted before.

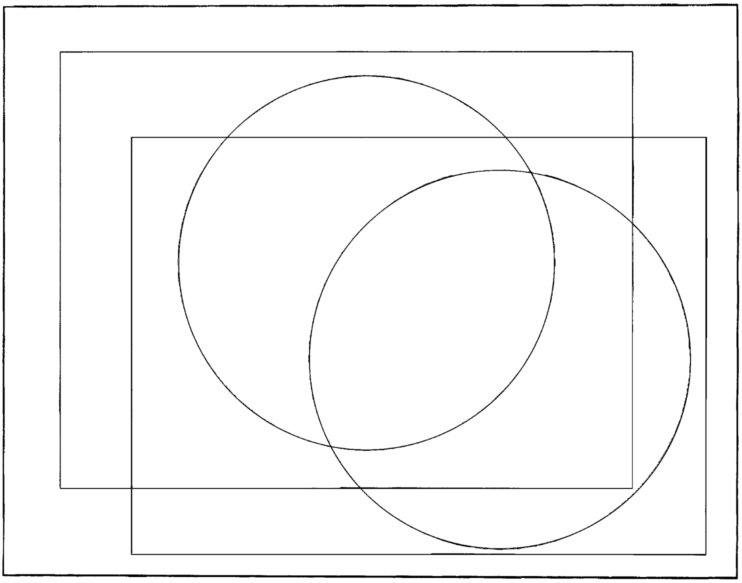
A problem that reviewers of the multisection validity literature face is that there appears no consensus, across student rating forms, of what constitutes the structure or dimensionality of instructional effectiveness as perceived by students. The effectiveness of postsecondary instruction is, like the elephant in *The Blind Men and the Elephant* (John Godfrey Saxe), a beast with many different characteristics. Each reviewer has attempted to examine this issue, and like the blind men of the poem, is convinced that he/she has discovered the true and accurate representation. What the following analysis attempts is to fit together the pieces to form a picture. But the analogy of the blind men and the elephant is not entirely correct. Each researcher does not hold only a unique piece of the puzzle but rather may hold some pieces in common with one or more other researchers, as wellas some unique pieces.

The question of the dimensionality or structure of instructional effectiveness across student rating forms can be approached in two ways: conceptually or empirically. In a conceptual or logical approach, theoretical models are used to develop a hierarchical structure or taxonomy. Borich (1977) suggests three stages in the development of a valid system of evaluating teacher effectiveness. The first stage is to search the literature for significant relationships and rationally select promising behaviors and skills. The second stage is to build a nomological network indicating antecedent, intervening and terminal behaviors, to test the validity of the above relationships, and to sequentially order the behaviors and skills. The third stage is to construct a taxonomy or hierarchy of behaviors emphasizing the important distinctions and minimizing the superfluous ones. Thus, the three stages are: selecting variables on the basis of the literature, chunking variables on the basis of relationships, and proposing higher-order structures on the basis of theory. The proposed hierarchical relationships among variables can then be empirically tested on a second sample via confirmatory factor analysis (CFA) or linear structured relationships (LISREL) (Hill, 1984).

There have been some studies attempting to elucidate empirically the structure of the student rating forms used in the multisection validity literature (Widlak et al., 1973; Kulik and McKeachie, 1975; Marsh, 1987, 1991). However, in general, one student rating form is factor analyzed and no attempt is made to compare it's factor structure to those of other rating forms purporting to measure the same dimensions of effective instruction. One exception is Marsh (1987) who commented on the similarity of specific factors in a number of student rating forms. However, this was done on the basis of a logical analysis and not empirically. Marsh (1991) analyzed logically the correspondence among the SEEQ, the Endeavor and Feldman's categories. He concluded that: Feldman's categories were much more specific than factors from either the SEEQ or the Endeavor; the SEEQ represented more of Feldman's categories than the Endeavor represented; and many SEEQ factors contained more than one Feldman category.

In the last few years, we have been using multivariate approaches to meta-analysis to explore the "common" factor structure across multiple student rating forms (Rosenfield, d'Apollonia, and Abrami, 1993; d'Apollonia, Abrami, and Rosenfield, 1993). Thus, we have combined both conceptual and empirical approaches. Figure 1 illustrates our goals. Take the large rectangle that surrounds the illustration. This represents all the qualities of teaching that could be represented in student rating forms. Any one rating form is represented by a smaller rectangle. Hence, two rating forms are illustrated in the figure. Each rating form rectangle is a subset of the whole. Furthermore, the rating form rectangles do not perfectly overlap, suggesting they represent somewhat different aspects of instruction. The circle within each rating form rectangle represents the rating form variability explained by a particular factor analysis of student responses to the rating form. Finally, there is an area of overlap between the circles representing the two rating forms. This is what is common to the factor analyses of the two rating forms. The non-intersecting part of the two circles represents what is unique to each of the factor analyses.

Figure 1: Hypothetical illustration of the underlying traits "common" to two rating forms.



Now, imagine a more complex figure with 17 rating form rectangles and their 17 circles within. Bits of intersecting circles represent what is common to two or more rating forms. But the union of all the circles represents the qualities of teaching represented by factors underlying all of the forms together.[3]

"COMMON" DIMENSIONS OF TEACHING

We decided to employ the Feldman coding scheme to further explore the dimensionality of student ratings of instruction and the validity of those dimensions to predict the products of instruction, particularly, student achievement. We found several difficulties with the scheme: a) lack of operational definitions for the categories (use of exemplars only); b) high intercoder agreement by us (Cohen's kappa = .93) but lower agreement (.60) with items categorized by Feldman; and c) internal inconsistencies including ambiguity, multidimensionality, and overlap among the categories. Using Feldman's coding scheme as the basis of our work, we decided to revise the scheme using the following principles:

1. The coding scheme should not be ambiguous. The categories used to code items should be clear, comprehensive and succinct. The categories should be of more or less equal breadth.
2. The bipolar values of a category should be contained within it; for example, clear and unclear presentations, authoritarian and participatory class management, etc.
3. Both the product and the process orientations to a teaching behavior should be in the same category; for example, the instructor presenting the subject as interesting and the students being interested in the subject.
4. Since global evaluations (course instructor, perceived learning) are included, the remaining categories should only include specific statements.

We defined our coding scheme based upon the 1,184 items collected from the student rating forms used in the *multisection validity studies*. Two coders subsequently coded the above items and obtained a 91.5%

[3] For methodological reasons associated with the aggregated correlation matrix, we were only able to examine that part of the union of the 17 circles that lay entirely encompassed within the largest circle.

intercoder agreement. The items from the *factor analytic studies* are a subset of the above items. The definition of each category is presented in Appendix A, beginning on page 253. The appendix also includes all items whose correlation were used to form the aggregate correlation matrix.

Collection of Factor Studies

We first collected studies that reported either complete factor matrices or correlation matrices for the student rating forms used in the multisection validity studies. We collected seventeen studies representing most of the rating forms in the validity set (excluding the in-house forms). One student rating form, the form used by Wherry (1951), supplied almost 50% of the items in the data set. It thus furnished a large portion of the interitem correlation coefficients. As expected, the global items are underrepresented in the factor set relative to the validity set.

Extraction and Coding of Outcomes

The outcome variables of interest for the integration of factor studies are the interitem correlation coefficients for each student rating form. These were estimated from the reproduced correlation matrices computed from the factor loading matrix of the items in the student rating forms (if rotated orthogonally), or from the pattern matrix and factor correlation matrix (if rotated obliquely). The 458 items from the factor studies were initially placed into 40 categories. These categories are listed and briefly defined in Appendix A.

Pruning and Synthesis of Aggregate Correlation Matrix

In order to aggregate the interitem correlation coefficients, one must first establish that the values being aggregated are homogeneous. If the set is not homogeneous, the (weighted) mean correlation does not properly represent the set of studies. There are a number of possible causes for heterogeneity: a) the items are ambiguous and/or multidimensional; b) the categories are ambiguous and/or multidimensional; and c) the relationship between items varies with setting, subject, etc.

The first two reasons speak to technical problems with the coding schema and certain student rating forms. Unfortunately, these problems confound questions concerning the dimensionality of effective instruction. Therefore, if one wishes to address the latter question, one must first reduce these technical problems. We therefore, eliminated (pruned) items and categories that were heterogeneous in the following manner.

We pruned items and categories from our data set in two stages. In the first stage we eliminated items that contributed to "poor" correlations between items belonging to the same category. We subdivided the complete set of interitem correlations (21,383 correlations) into 40 sets of interitem correlations between items belonging to the same category. We assumed that if the categories were unidimensional and generalizable, sets of interitem correlation coefficients should be uniform across student rating forms and the mean interitem correlation coefficient for the set should approach 1.0. In other words, the mean interitem correlation coefficient for the subset is analogous to a reliability coefficient. For each set, we identified items that contributed to correlations that were below 0.5, or that lowered the mean interitem correlation coefficient for a category consistently below .65. We scrutinized these items for ambiguous wording, reversed polarity, negative wording, compound statements, etc. We subsequently dropped these items. For each set, we continued pruning until the set was homogeneous (*i.e.* the coefficient of variability was .20 or less). We eliminated three categories, *appropriate use of materials, low-level cognitive outcomes*, and *overall learning* because of insufficient data.

In the second stage we eliminated items that contributed to heterogeneous correlations between items in different categories. This is a more difficult task in that since the correlations for items belonging to different categories are not known we can no longer assume that the correlations should approach 1. However, we can still expect that there should be a tight cluster of values about the weighted mean.

We subsequently subdivided the remaining 8,131 correlations into 666 sets representing the intercorrelations between items belonging to different categories. Taking one set at a time, we identified the items that contributed to correlations at the extremes of the distribution. We eliminated those items that consistently contributed to the heterogeneity of the set. Finally, after all pruning had been done, we reviewed all decisions to see if later decisions to drop items would allow us to reinsert some dropped items. Note that two items contributed to a correlation that was an "outlier." In some cases the "poor" item could be easily identified because of poor wording, double negatives, compound items, etc. However in other cases, the choice of item to be eliminated was somewhat arbitrary. That is, there is not one unique set of items which if eliminated produce homogeneous sets. Rather, there are a number of possible sets. Moreover, interitem correlations exist only between items in the same student rating form. Therefore the distribution of items per category/per rating form influences which items can be considered for elimination. That is, there have to be at least two items within a category from the same rating form for either item to be considered for pruning at the first stage. Therefore, some of the items that were retained at both pruning stages were retained not because they are "superior" items, but rather because they were never considered for elimination.

In addition, we eliminated two more categories, *time management* and *workload*, because we were not able to reduce the heterogeneity without deleting all the items in some sets. Less than 2% of the 595 sets that remained are heterogeneous. We decided not to drop any other categories or items since they did not consistently produce heterogeneity across all sets.

Thus, we constructed a 35 by 35 correlation matrix. This matrix represented the aggregation of 6,788 interitem correlations computed for 225 items from 17 rating forms. These items, sorted by category, are presented in Appendix A.

Of the 40 Feldman instructional categories, only five were missing from this correlation matrix because of either excessive heterogeneity or insufficient data. These five categories were: *appropriate use of methods/materials, low-level cognitive outcomes, overall learning, time management* and *workload.*

Factor Analysis

Factors were extracted from the aggregate correlation matrix produced above using SPSS (SPSS Inc., 1990). Factors with eigenvalues greater than 1.0 were extracted. The solution was then rotated obliquely using OBLIMIN with a delta of .2. Four factors were extracted via principal components extraction. The percent variance extracted by each factor, in decreasing magnitude, were 62.8%, 4.2%, 3.7, and 2.9%. Of the 35 categories, all except *course objectives, knowledge of domain*, and *supervision and disciplinary activities* had loadings of at least .62 on the first component. Thus there clearly is a large general factor which explains about 63% of the variance in student ratings.

In order to improve interpretability, the solution was rotated obliquely; thus, the variance was redistributed over the four factors.

The four factors, in order of importance as judged by the sum of squared loadings, are described below.

Thirteen categories load on Factor 1 (loadings > .55): *choice of supplementary materials, relevance of instruction, overall course, monitoring learning, general knowledge and cultural attainment, research productivity and reputation, motivating students to greater effort, enthusiasm for teaching, high-level cognitive outcomes, clarity of instruction, stimulation of interest, preparation*, and *management style*. We note that most of these categories pertain to the instructor viewed in an instructional role. The sum of the squared loadings is 8.0 and, therefore, this factor appears to be the most important factor in instructional effectiveness.

Sixteen categories load on factor 2 (loadings > .38): *personal appearance, health, and attire, general attitudes, dramatic delivery, concern for students, vocal delivery, answering questions, knowledge of teaching, tolerance of diversity, availability, overall instructor, interaction and discussion, respect for others, enthusiasm for students, friendly classroom climate, enthusiasm for subject*, and *personality characteristics*. We note that most of these categories pertain to the instructor viewed as a person. The sum of the squared loadings is 5.6 and, therefore, the second factor is almost as important as the first factor.

Two categories load on Factor 3 (loadings > .75): *evaluation* and *feedback*. We note that these two categories pertain to the instructor viewed as a regulator. The sum of the squared loadings is 2.5 and, therefore, this factor is considerably less important than the previous two.

Four categories load on Factor 4. These are *supervision and disciplinary actions, knowledge of domain, choice of required materials*, and *objectives*. The sum of the squared loadings is 1.8 and, therefore, this factor is the least important factor. We note that it is difficult to interpret this factor, but it is considerably less important than the previous three and may not be stable. It is also the only factor that is not correlated with the other factors.

Since we aggregated items within categories and factor analyzed relatively homogeneous categories, one would expect to extract "higher-order" factors representing the "common" aspects of instruction across situations. We extracted 62.8% of the variance across the interitem correlations in the first principal component. All items load heavily on this component, suggesting it is an overall instructional skill factor. Such a general skill factor has been proposed by Kulik and McKeachie (1975). Rotation results in the redistribution of variance such that three correlated factors emerged, along with one subsidiary uncorrelated factor. These three correlated factors are similar to the three factors proposed by Widlak, McDaniel, and Feldhusen (1973) describing three roles: instructor, actor, and director. They subsequently factor analyzed responses to the 18 item Course-Instructor-Evaluation form from Purdue and obtained three highly correlated factors.

Feldman (1976) also investigated the pattern of relationships among factors from 60 factor studies. He reported that "despite the profusion of connections…, a fairly consistent and meaningful pattern does emerge: indeed, this pattern supports the view of Widlak et al. (1973) that instructors primarily enact three different roles." Feldman called these roles presentation, facilitation, and regulation.

To address concerns related to both the number of items pruned and the possibility of alternative pruned sets, we ran similar factor analyses of the complete data set, and alternative pruned sets. In all cases the same general factor emerged. Differences occurred primarily in those categories having moderate loadings (e.g., those mentioned above as loading on two factors).

Our factor analysis across the multiple rating forms indicates that there is a "common" structure to instructional effectiveness. Four factors were obtained, three of which were highly correlated. Global items were loaded highly on the first two factors.

The finding that the factors were correlated may obscure setting differences. For example, some students (e.g., engineering students in calculus classes) may respond favorably to the clarity of instruction and give especially high mean ratings for clarity, while other students (e.g., psychology students in clinical classes) may respond favorably to an interactive classroom climate and give especially high mean ratings for interaction and discussion. Despite situational differences in mean ratings, a "common" correlated factor structure would emerge.

Whatever the reason for the high correlations between the factors, the finding that there is such a large global component and that it is highly correlated with the other components argues against the utility of using specific factors or teaching categories to make summative assessments of instruction. We believe that the logical and empirical analyses already presented by us and others provide support for a large, underlying general trait "effective teaching" although it may not be the only trait. In addition, we believe that effective teaching is multidimensional but that there is inconsistency concerning the teaching dimensions, particularly across rating forms (i.e., operationalizations of different latent traits). This inconsistency suggests that any one of the existing multidimensional rating forms may not represent teaching for all instructors, courses, and settings. Therefore, we recommend that specific ratings should be used cautiously for summative decisions about teaching. If one uses an existing rating form, computing a composite score based on the categories and items that formed the general factor of our analysis would appear to be superior to using separate dimensional scores. If one prepares a customized form, only those items and categories that loaded highly on the first principal component of our analysis should be included and their scores averaged. Finally, it remains our opinion that the best alternative to averaging across specific items is to base summative decisions of teaching effectiveness on global ratings.

CONCLUSIONS

Numerous studies have explored the dimensions of effective college instruction. Yet there remain notable disagreements regarding whether and how data from multidimensional student rating forms should be used in summative decisions about teaching. This paper critically examined a host of issues associated with the dimensions of instructional effectiveness as reflected in student ratings of teaching. We discussed effective teaching from both product and process views. From the product view, effective teaching can be defined as the positive cognitive, affective, and/or psychomotor changes produced in students. From the process view, effective teaching can be defined as the teaching activities that occur both before (preparatory) and during (delivery) teaching. We subsequently discussed the need for research exploring the impact of process variables on product variables. The second section provided a general discussion of methods for empirically determining effective teaching. The third section concentrated on the strengths and weaknesses of three validation designs—the laboratory design, the multisection design and the multitrait-multimethod design. The fourth section summarized the quantitative literature reviews of the 43 multisection validity studies. Finally, the fifth section considered factor analysis and the dimensions of effective teaching. We summarized our attempts to quantitatively integrate the results from seventeen correlation matrices by coding the items using a common scoring scheme, eliminating items that were heterogeneous within categories, and factor analyzing the aggregated correlation matrix.

We conclude that existing analyses provide support for a large underlying general trait although it may not be the only trait. We also believe that effective teaching is multidimensional but that there are differences across rating forms concerning the specific dimensions that underlie effective instruction. These differences suggest that student ratings of specific teaching dimensions should not be used indiscriminately for summative decisions about teaching effectiveness.

In this paper we have presented many lines of evidence that suggest that although instructional effectiveness is multidimensional, global items should be used for the purposes of summative decisions. First, when examining many rating forms one is immediately struck by the fact that, despite their differences, what they share is a similar set of global items. Second, global items, more so than many specific instructional dimensions, have relatively high validity coefficients. Third, different instructional settings are likely to have larger effects on specific dimensions than on global items. Fourth, even in well designed multidimensional forms, such as the SEEQ, global items load most strongly on the first few factors. Finally, our factor analysis across seventeen rating forms confirms the four points listed above.

Appendix A: Student Rating Items and Their Categories

In the list below, categories are arranged alphabetically. The five categories not used in the final analysis are presented solely for completeness and are marked with an asterisk (*). These categories are listed with definitions while all other categories contain definitions as well as all items retained for final analysis. Note that items are presented as in original sources. If an item appeared in multiple sources it presented multiply. A quadruple of numbers appears immediately after the name of the category (used in the final analysis) to represent: a) the initial number of items code; b) the number of items retained through all stages; c) the number of items dropped at the stage when interitem correlations **within** each category were examined; and d) the number of items dropped at the stage when interitem correlations **between** categories were examined).

Answering Questions (9/6/2/1): The students are evaluating the extent to which the instructor encouraged students to ask questions and responded to students' questions appropriately.

Rate the instructor on the basis that he answers student's questions in a clear and concise manner.
Rate the extent to which the instructor responded effectively to student questions.
Encouraged questions and answers.
The instructor encouraged and readily responded to student questions.
Became angry when questions were asked.
No questions allowed between explanations.

*__Appropriate Use of Methods/Materials (2/2/0/0)__: The students are evaluating the extent to which the instructor uses appropriate instructional methods and materials in class, including appropriate use of textbook and tests for learning.

Availability (7/4/0/3): The students are evaluating the extent to which the instructor was available outside of the classroom for assistance or extra-curricular activities.

> Rate the instructor on the basis of the ease at which an office appointment can be made.
> Welcomed seeking help and advice.
> Accessible to individual students.
> Welcomed conferences.

Choice of Required Materials (8/4/1/3): The students are evaluating the qualities of the required course materials including textbooks, assignments, etc.

> The textbook was very good.
> Readings and text valuable.
> Assignments added to course understanding.
> Did not go to trouble of making up assignments.

Choice of Supplementary Materials (4/2/0/2): The students are evaluating the qualities of the supplementary materials (e.g., film, audio-visuals, etc.). That is, they are evaluating whether they were interesting, valuable, or personally relevant. Unless explicitly labeled "supplementary" such materials are considered to be required.

> The outside assignments for this course are just about the right length/somewhat too long/somewhat too short/much too long/much too short.
> Had varied illustrations about topic covered.

Clarity of Instruction (25/15/3/7): The students are evaluating the extent to which the instructor delivers clear, concise, understandable and accurate instruction (e.g., lectures, laboratories, etc.).

> Presentation of subject matter.
> Rate the instructor on the basis of the organized class presentation.
> Rate the instructor on the basis that she makes clear or simple the difficult ideas or concepts in this course.
> The instructor did not synthesize ideas.
> Rate the extent to which the instructor was successful in explaining the course material.
> Presentations clarified material.
> Presented clearly and summarized.
> Instructor's explanations clear.
> Presentation well prepared and integrated.
> He explained clearly and his explanations were to the point.
> Instructions not complete.
> Covered subject well.
> Made subject clear.
> Presentations of materials especially good.
> Students in constant state of uncertainty.

Concern for Students (8/6/0/2): The students are evaluating the extent to which the instructor was concerned and helpful about student difficulties

The instructor seemed genuinely concerned with student's progress and was actively helpful.
The instructor seemed to be concerned with whether the students learned the material.
Listened and willing to help.
Concerned about student difficulties.
The instructor maintained a generally helpful attitude toward students and their problems.
Too busy for talks with students.

Dramatic Delivery (5/3/2/0): The students are evaluating the extent to which the instructor delivered instruction in an expressive, dynamic, dramatic or exaggerated manner.

Dynamic and energetic.
Talked with back to class.
Hard to believe.

Enthusiasm for Students (11/9/2/0): The students are evaluating the extent to which the instructor communicates his/her enthusiasm, interest or liking for students as people.

Sympathetic attitude toward students.
Rate the instructor on the basis of the instructor's apparent interest in working with students.
The instructor seemed to be interested in students as persons.
Interested in individual students.
Was the instructor considerate of and interested in his students?
Always suspicious of students.
Afraid of students.
Lacked interest in students.
Kept up with student affairs.

Enthusiasm for Subject (3/3/0/0): The students are evaluating the extent to which the instructor communicates his/her enthusiasm, interest or liking for the subject.

Interest in subject.
The instructor was enthusiastic when presenting course material.
Interested in all aspects of subject.

Enthusiasm for Teaching (4/3/0/1): The students are evaluating the extent to which the instructor communicates his/her enthusiasm, interest or liking for teaching.

The instructor seemed to consider teaching as a chore or routine activity.
Enthusiastic about teaching.
Enjoyed teaching class.

Evaluation (27/8/11/8): The students are evaluating the extent to which the instructor's tests were appropriate in terms of content, frequency, time allocation, weight, difficulty, validity and learning opportunity. They are also evaluating the instructor's fairness and consistency in grading.

> The types of test questions used were good.
> Fair and impartial grading.
> Grading reflected performance.
> Grading indicated accomplishments.
> Evaluation methods fair and appropriate.
> Exams emphasized course content.
> Tests indicated careful preparation.
> Would not explain grading system.

Feedback (16/5/8/3): The students are evaluating the instructor's use of review and feedback (frequency, positive/negative) and its effect on students.

> Instructor did not review promptly and in such a way that students could understand their weaknesses.
> The instructor made helpful comments on papers or exams.
> Rate the instructor on the basis of the information or feedback provided concerning the nature and quality of my work (considering all the factors involved in teaching this course).
> Examination feedback valuable.
> Reviewed test questions that majority of students missed.

Friendly Classroom Climate (8/6/2/0): The students are evaluating the extent to which the instructor modeled, encouraged and achieved a friendly and safe classroom.

> He was friendly.
> Friendly towards students.
> Discouraged students.
> Made students feel very insecure.
> Very much at ease with the class.
> Students often returned to chat with teacher.

General Attitudes (4/3/1/0): The students are evaluating the instructor's general attitudes. (An attempt is first made to fit items into the other, more specific instructional dimensions. Only if they do not fit elsewhere are they classified here.)

> Liberal and progressive attitude.
> Had unethical attitudes.
> Did not approve of extracurricular activities.

General Knowledge and Cultural Attainment (2/2/0/0): The students are evaluating the instructor's general knowledge and cultural attainment beyond the course.

> Admired for great intelligence.
> Large background of experience made subject more interesting.

High-level Cognitive Outcomes (32/11/21/0): The students are evaluating the extent to which the instructor is promoting high-level cognitive outcomes such as writing skills, reasoning, meta-cognition, problem solving, etc.

The instructor encouraged students to think for themselves.
The instructor encouraged the development of new viewpoints and appreciations.
Understand advanced material.
Ability to analyze issues.
I can think more coherently.
Developing a sense of personal responsibility (self-reliance, self-discipline).
Discovering the implications of the course material for understanding myself (interests, talents, values, etc.).
Developing specific skills, competencies and points of view that I can use later in life.
Intellectual curiosity in subject stimulated.
Gained general understanding of topic.
Encouraged students to think out answers.

Interaction and Discussion (15/6/1/8): The students are evaluating the extent to which the instructor modeled, encouraged and achieved interactive classes in which both students and instructor contributed to the class.

Encouraged class discussions.
Encouraged expression of ideas.
Students would not cooperate in class.
Group discussions encouraged.
Nothing accomplished in classroom discussions.
Very skillful in directing discussion.

Knowledge of Domain (4/1/2/1): The students are assessing the instructor's knowledge of the specific course subject matter and its applications.

Did not need notes.

Knowledge of Teaching and of Students (1/1/0/0): The students are evaluating the instructor's knowledge of pedagogy (e.g., knowledge of students, student learning, and/or of instructional methods).

No ability to handle students.

***Low-level Cognitive Outcomes**: The students are evaluating the extent to which the instructor is promoting low-level cognitive outcomes (e.g., recall, recognition, knowledge, etc.).

Management Style (23/10/12/1): The students are evaluating the instructor's management style (e.g., authoritarian/participatory, formal/informal) and method of handling issues of classroom control (e.g., noise, order, seating, calling on students).

The demands of the students were not considered by the instructor.
He decided in detail what should be done and how it should be done.
He was permissive and flexible.

Knack in dealing with all types of problems.
Never deliberately forced own decisions on class.
Classes always orderly.
Conducted class smoothly.
Never considered what class wanted.
Maintained a well organized classroom.
Weak in leadership questions.

Monitoring Learning (7/5/1/1): The students are evaluating the extent to which the instructor monitored students' reactions and taught at the appropriate individual and class level.

The instructor was skilful in observing student reactions.
Skilled at bringing out special abilities of students.
Worked with students individually.
Aware of individual differences in pupils.
Sensed when students needed help.

Motivating Students to Greater Effort (18/9/3/6): The students are evaluating the extent to which the instructor motivated students to more effort, intellectual curiosity, love of learning, high academic aspirations, etc.

Stimulating intellectual curiosity.
Rate the instructor on the basis that the teaching methods inspire, stimulate or excite me intellectually.
Rate the instructor on the basis that she motivates me to think rather than just memorize material.
I developed motivation to do my best work.
Plan to take more courses.
Inspired many students to do better work.
Motivated students to work.
Instilled spirit of research.
Inspired class to learn.

Objectives (11/4/3/4): The students are evaluating the extent to which the instructor communicated performance criteria and deadlines for assignments and tests.

The direction of the course was adequately outlined.
Detailed course schedule.
The instructor was clear on what was expected regarding course requirements, assignments, exams, etc.
Students always knew what was coming up next day.

Overall Course (8/5/2/1): The students are evaluating the overall worth and quality of the course.

You generally enjoyed going to class.
Overall course rating.
How would you rate the overall value of this course?
Have you enjoyed taking this course?
Students discouraged with course.

Overall Instructor (13/12/1/0): The students are evaluating the overall effectiveness of the instructor.

Rate the overall teacher's effectiveness.
General teaching ability.
Attitudes about teaching.
Would you recommend this course from this instructor?
Overall instructor rating.
Would you recommend this course from this instructor?
How would you rate your instructor with respect to general (all-around) teaching ability?
Overall evaluation of instructor.
Would like instructor as personal friend.
Learned a lot from teacher.
Students avoided this teacher's class.
Not qualified as a teacher.

***Overall Learning**: The students are evaluating the overall quality and relevance of the perceived learning that took place including the achievement of short and long term objectives.

Personal Appearance, Health, and Attire (11/5/6/0): The students are evaluating the instructor's personal appearance, health and attire.

Personal appearance.
Teacher very careless about dress.
Very pleasing appearance.
Wore wrinkled clothes.
Poor posture.

Personality Characteristics and Peculiarities (24/20/4/0): The students are evaluating the instructor's general personality characteristics and peculiarities not directly related to teaching (e.g., maturity, irritability, confidence, paranoia, cynicism, etc.).

Sense of proportion and humor.
Personal peculiarities.
Rate the instructor on the basis of poise and classroom mannerisms.
The instructor exhibited professional dignity and bearing in the classroom.
Enhanced presentations with humor.
Crabby.
Good natured.
Consistent.
A typical old maid (or bachelor) personality.
Immature emotionally.
Very prejudiced.
Considerate.
No sense of humor.
Tactless.
Wonderful sense of humor.
Cynical attitude repels students.

Did not inspire confidence.
Magnetic personality.
Tried to show off.
Well-rounded personality.

Preparation and Organization (13/8/2/3): The students are evaluating the extent to which the instructor prepared himself/herself for instruction. (This category only related to preparation, not presentation. Any items that are ambiguous in terms of whether they relate to preparation or presentation are classified as presentation are classified as presentation are classified as presentation since students judge on the basis of presentation.)

Course material was poorly organized.
Generally the course was well organized.
Rate the extent to which the instructor's lectures were well prepared.
The instructor was consistently prepared for class.
Rate the extent to which the instructor's lectures and other material were well prepared.
Absolutely no previous preparation for class.
Became confused in class.
Best organized of any class I have had.

Relevance of Instruction (11/7/3/1): The students are evaluating the extent to which the instructor emphasizes the relevance of the provided information, including recent research.

The instructor's use of examples or personal experiences helped to get points across in class.
Good use of examples.
Contrasted implications.
Gave background of ideas and concepts.
Gave different points of view.
Discussed current developments.
Related subject to everyday life.

Research Productivity and Reputation (3/2/1/0): The students are evaluating the instructor's research productivity and reputation.

Cooperative with other teachers.
Looked to for advice.

Respect for Others (28/15/13/0): The students are evaluating the extent to which the instructor modeled, encouraged and showed trust, respect, and consideration for others (e.g., listened without interruption, did not not belittle or criticize others' criticism, treated others as equals, was punctual, etc.).

The instructor's attendance and punctuality have been consistently good.
He listened attentively to what class members had to say.
Irritated easily.
Very impatient with less able students.
Carried friendliness outside of classroom.

Built up confidence in students.
Gained class confidence very quickly.
Made students feel at ease.
Sarcastic if disagreed with.
Students did things to make teacher mad.
Always very polite to students.
Humiliated students.
Publicly ridiculed some students.
Ridiculed students.
Very sincere when talking to students.

Stimulation of Interest in the Course (21/13/4/4): The students are evaluating the extent to which the instructor stimulated their interest in the course by using a variety of activities, manifested by the extent to which good attendance, increased interest, outside reading, and liking/enjoyment for the subject matter were exhibited.

Rate the instructor on the basis that she presents the material or content of this course in an interesting manner.
Rate the extent to which the instructor stimulated your interest in the course.
Increased subject interest.
Teaching style held your interest.
Rate the extent to which the instructor stimulated your interest in the course.
Do you now enjoy reading more than you used to?
Gained interest in American government.
Do more reading on topic.
Everyone attended regularly.
Knew how to hold attention in presenting materials.
Made lectures stimulating.
No attempt to make course interesting.
Students counted the minutes until class was dismissed.

Supervision and Disciplinary Actions (3/1/2/0): The students are evaluating the extent to which the instructor supervised tests and handled disciplinary actions when disruptions occurred.

Never had to discipline the students.

*__Time Management__: The students are evaluating the extent to which the instructor handled class time.

Tolerance of Diversity (12/6/3/3): The students are evaluating the extent to which the instructor modeled, encouraged and achieved tolerance for a diversity of opinions, ideas and viewpoints and an absence of prejudice in the classroom.

The instructor was open to other viewpoints.
Rate the instructor on the basis that he considers opposing viewpoints or ideas.
The instructor appeared receptive to new ideas and others' viewpoints.
Intolerant.
Presented both sides of every question.
Blinded to all viewpoints but own.

Vocal Delivery (7/5/2/0): The extent to which the instructor demonstrated skill in vocal delivery.

> Rate the instructor on the basis that she speaks clearly and is easily heard.
> The instructor is clear and audible.
> Speech very fluent.
> Lectured inaudibly.
> Occasional bad grammar detracted from speech.

*****Workload**: The students are evaluating the performance standards and the workload (amount, difficulty) of the course and assignments.

References

Abrami, P.C. (1984, February). Using meta-analytic techniques to review the instructional evaluation literature. *Postsecondary Education Newsletter* 6: 8.

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. (1989a). SEEQing the truth about student ratings of instruction. *Educational Researcher* 43(1): 43–45.

Abrami, P.C. (1989b). How should we use student ratings to evaluate teaching? *Research in Higher Education* 30: 221–227.

Abrami, P.C., Chambers, B., Poulsen, C., DeSimone, C., d'Apollonia, S., and Howden, J. (1995). *Classroom Connections: Understanding and Using Cooperative Learning*. Toronto, Ontario: Harcourt Brace.

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

Abrami, P.C., and d'Apollonia, S. (1987, April). *A Conceptual Critique of Meta-analysis: The Literature on Student Ratings of Instruction*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Abrami, P.C., and d'Apollonia, S. (1988, April). *The Literature on Student Ratings of Instruction: A Conceptual Solution to Some Implementation Problems of Meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Abrami, P.C., and d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall and J. Franklin (eds.) *Student Ratings of Instruction: Issues for Improving Practice. New Directions for Teaching and Learning*. Number 43, pp. 97–111. San Francisco: Jossey-Bass.

Abrami, P.C., and d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness-Generalizability of "N = 1" research: Comment on Marsh (1991). *Journal of Educational Psychology* 83: 411–415.

Abrami, P.C., d'Apollonia, S., and Cohen, P.A. (1990). The validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology* 82: 219–231.

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

Abrami, P.C., and Mizener, D.A. (1985). Student/instructor attitude similarity, student ratings, and course performance. *Journal of Educational Psychology* 77: 693–702.

Anderson, N.H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology* 9: 272–279.

Borich, G.D. (1977). *The Appraisal of Teaching: Concepts and Process*. Reading, MA: Addison-Wesley

Bushman, B.J., Cooper, H.M., and Lemke, K.M. (1991). Meta-analysis of factor analyses: An illustration using the Buss-Durke Hostility Inventory. *Personality and Social Psychology Bulletin* 17: 344–349.

Bloom, B.S., Engelhart, M.D., Frost, E.J., Hill, W.H., and Krathwohl, D.R. (1956). *Taxonomy of Educational Objectives. Handbook I: Cognitive Domain*. New York: David McKay.

Campbell, D.T., and Stanley, J.C. (1966). *Experimental and Quasi-experimental Designs for Research*. Boston: Houghton-Mifflin.

Cashin, W.E., and Downey, R.G. (1992). Using global student rating items for summative evaluations. *Journal of Educational Psychology* 84: 563–572.

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.

Cohen, P.A. (1982). Validity of student ratings in psychology courses: A meta-analysis of multisection validity studies. *Teaching of Psychology* 9: 78–82.

Cohen, P.A. (1983). Comment on a selective review of the validity of student ratings of teaching. *Journal of Higher Education* 54: 448–458.

Cohen, P.A. (1986, April). *An Updated and Expanded Meta-analysis of Multisection Student Rating Validity Studies*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. CA.

Cohen, P.A. (1987, April). *A Critical Analysis and Reanalysis of the Multisection Validity Meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Cooper, J., Prescott, S., Cook, L., Smith, L., Mueck, R., and Cuseo, J. (1990). *Cooperative learning and College Instruction: Effective Use of Student Learning Teams*. Long Beach, CA: California State University Foundation on behalf of California State University Institute for Teaching and Learning, Office of the Chancellor.

d'Apollonia, S. and Abrami, P.C. (1987, April). *An Empirical Critique of Meta-analysis: The Literature on Student Ratings of Instruction*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

d'Apollonia, S. and Abrami, P.C. (1988, April). *The Literature on Student Ratings of Instruction: Yet Another Meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

d'Apollonia, S., Abrami, P., and Rosenfield, S. (1993, April). *The Dimensionality of Student Ratings of Instruction: A Meta-analysis of the Factor Studies*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

d'Apollonia, S., Abrami, P., and Rosenfield, S. (in preparation). A multivariate meta-analysis of the multisection validity studies.

Doyle, K.O., Jr. (1981). Validity and perplexity: An incomplete list of disturbing issues. *Instructional Evaluation*, 6(1): 23–25.

Doyle, K.O., and Crichton, L.I. (1978). Student, peer, and self-evaluations of college instructors. *Journal of Educational Psychology* 5: 815–826.

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

Feldman, K.A. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education* 6: 223–274.

Feldman, K.A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education* 18: 3–214.

Feldman, K.A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education* 21: 45–116.

Feldman, K.A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher Education* 28: 291–344.

Feldman, K.A. (1989a). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education* 30: 583–645.

Feldman, K.A. (1989b). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education* 30: 137–194.

Feldman, K.A. (1990). An afterword for "The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the syn<??>hesis of data from multisection validity studies." *Research in Higher Education* 31: 315–318.

Fernald, P.S. (1990). Students' ratings of instruction: Standardized and customized. *Teaching of Psychology* 17: 105–109.

Franklin, J., and Theall, M. (1989, April). *Rating the Readers: Knowledge, Attitude, and Practice of Users of Student Ratings of Instruction*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

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

Gaski, J.F. (1987). On "Construct validity of measures of college teaching effectiveness." *Journal of Educational Psychology* 79: 326–330.

Gorsuch, R.I. (1983). *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Hedges, L.V., and Olkin, I. (1985). *Statistical Methods for Meta-analysis*. Orlando, FL: Academic Press.

Hill, P.W. (1984). Testing hierarchy in educational taxonomies: A theoretical and empirical investigation. *Evaluation Education* 8: 181–278.

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

Isaacson, R.L., McKeachie, W.J., Milholland, J.E., Lin, Y.G., Hofeller, M., Baerwaldt, J.W., and Zinn, K.L. (1964). Dimensions of student evaluations of teaching. *Journal of Educational Psychology* 55: 344–351.

Johnson, D.W., Johnson, R.T., and Smith, K.A. (1991). *Active Learning: Cooperation in the College Classroom*. Edina, MN: Interaction Book Co.

Kaiser, H.F., Hunka, S., and Bianchini, J.C. (1969). *Relating Factors Between Studies Based Upon Different Individuals*. In H.J. Eysenck (ed.), Personality Structure and Measurement. San Diego, CA: Knapp.

Kulik, J.A., and McKeachie, W.J. (1975). The evaluation of teachers in higher education. *Review of Research in Education* 3: 210–240.

Linn, R.L., Centra, J.A., and Tucker, L. (1975). Between, within, and total group factor analysis of student ratings of instruction. *Multivariate Behavioral Research* 10: 277–288.

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

Marsh, H.W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology* 83: 285–296.

Marsh, H.W. (1991). A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami and d'Apollonia. *Journal of Educational Psychology* 83: 416–421.

Marsh, H.W., and Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. Smart (ed.), *Higher Education: Handbook of Theory and Research*, Vol. VIII. New York: Agathon Press.

Marsh, H.W., and Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal* 21: 341–366.

Marsh, H.W., and Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education* 7: 9–18.

Maxwell, S.E., and Howard, G.S. (1987). On the underdetermination of theory by evidence. *Journal of Educational Psychology* 79: 331–332.

McCallum, L.W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education* 21: 150–158.

Murray, H.G. (1991). Effective Teaching Behaviors in the College Classroom. In J. Smart (ed.), *Higher Education, Handbook of Theory and Research* (Vol. 6) New York: Agathon.

Murray, H.G., Rushton, J.P., and Paunonen, S.V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology* 82: 250–261.

Osgood, C.E., Suci, G.J., and Tannenbaum, P.H. (1957). *The Measurement of Meaning*. Urbana, IL: University of Illinois Press.

Rosenfield, S., d'Apollonia, S., and Abrami, P.C. (1993, April). *The Dimensionality of Student Ratings of Instruction: Aggregating Factor Studies*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Seldin, P. (1991). *The Teaching Portfolio*. Bolton, MA: Anker Publishing Co.

Shore, B.M., Foster, S.F., Knapper, C.K., Nadeau, C.G., Neill, N., and Sim, V.W. (1986). *The Teaching Dossier: A Guide to Its Preparation and Use*. Ottawa: Canadian Association of University Teachers.

Smith, R.A., and Cranton, P.A. (1992). Students' perceptions of teaching skills and overall effectiveness across instructional settings. *Research in Higher Education* 33: 747.

Sullivan, A.M., and Skanes, G.R. (1974). Validity of student evaluations of teaching and the characteristics of successful instructors. *Journal of Educational Psychology* 66: 584–590.

Tabachnick, B.G., and Fidell, L.S. (1983). *Using Multivariate Statistics*. New York: Harper and Row.

Thomson, B. (1989). Meta-analysis of factor structure studies: A case study example with Bem's Androgyny measure. *Journal of Experimental Education* 57: 182, 197.

Wherry, R.L. (1951). *The Control of Bias in Ratings: Factor Analysis of Rating Item Content*. Columbus: The Ohio State University Research Foundation, United States Army, AGO, Personnel Research Branch, PRB Report No. 919.

Widlak, F.W., McDaniel, E.D., and Feldhusen, J.F. (1973). *Factor Analyses of an Instructor Rating Scale*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service ED 079324).

The Dimensionality of Student Ratings of Instruction: An Update on What We Know, Do not Know, and Need to Do

Philip C. Abrami*, **Steven Rosenfield**† **and Helena Dedic**†

**Concordia University*

abrami@education.concordia.ca

†*Vanier College*

Abstract

Analysis ten years ago of seventeen multidimensional student rating forms revealed that all included global items measuring one underlying general trait: effective teaching. Differences existed across rating forms concerning which other underlying dimensions were included, which led to the conclusion that only student ratings of global items, not those concerning specific teaching dimensions, should be used for summative decisions about teaching effectiveness. In response to society's current transition into the Information and Communication Age, a transformation of learning environments has begun, without parallel changes in rating forms. Thus, summative use of dimensions other than the general trait of effective teaching, now even less relevant to, perhaps even biased against, teachers transforming their learning environments, may be slowing educational reform.

Key Words: student ratings, effective teaching, post-secondary education, science education, metanalysis

In our earlier article for *Higher Education: A Handbook of Teaching and Learning*, Abrami, d'Apollonia, and Rosenfield (1996) explored the dimensionality of instruction as reflected in student ratings. While ten years does not seem like a terribly long time in the social sciences, we worked on this update wondering whether our conclusions stood the test of time.

Abrami et al. (1996) was divided into five sections. In the first section, we explored three alternative definitions of effective teaching.

* Address inquires to: Philip C. Abrami, Centre for the Study of Learning and Performance, Concordia University, 1455 DeMaisonneuve Blvd. W., Montreal, Quebec Canada H3G 1M8. E-mail: abrami@education.concordia.ca. Website: http://doe.concordia.ca/cslp/. Support for the preparation of this article was provided by grants to the authors from Fonds québécois de la recherche sur la société et la culture, Government of Quebec. The authors are solely responsible for the content of the article.

In section two, we discussed methods of empirically determining effective teaching. In section three, we concentrated on the strength and weaknesses of student ratings validation designs. In section four, we examined the quantitative reviews of the validation studies. And in section five, we summarized our analyses of the common dimensions of effective teaching as reflected in student ratings by integrating the results of seventeen correlation matrices.

We concluded "existing analyses provide support for a large underlying general trait although it may not be the only trait. We also believe that effective teaching is multidimensional but that there are differences across rating forms concerning the specific dimensions that underlie effective teaching. These differences suggest that student ratings of specific teaching dimensions should not be used indiscriminately for summative decisions about teaching effectiveness.

In this paper we have presented many lines of evidence that suggest that although instructional effectiveness is multidimensional, global items should be used for the purposes of summative decisions. First, when examining many rating forms one is immediately struck by the fact that, despite these differences, what they share is a similar set of global items. Second, global items, more so than many specific instructional dimensions, have relatively high validity coefficients. Third, different instructional settings, involving disciplinary differences, year, career path, etc. are likely to have larger effects on specific dimensions than on global items. Fourth, even in well designed multidimensional forms, such as the SEEQ, global items load most strongly on the first few factors. Finally, our factor analysis across seventeen rating forms confirms the four points listed above" (Abrami et al., 1996, p. 357).

In preparing the current update, we focus on two aspects of student ratings that build upon our earlier findings. The first concerns a contrast between student-centred and teacher-centred learning environments and how student ratings of specific teaching qualities should vary across these environments. The second concerns the use of ratings for summative decisions about teaching effectiveness.

STUDENT-CENTRED AND TEACHER-CENTRED LEARNING ENVIRONMENTS

In this section we will discuss the results of a study (Rosenfield et al., 2005) of learning environments in post-secondary science classrooms. This study was undertaken in large part because reports indicate a looming shortage of graduates in science and engineering from our universities in North America (OECD, 2005; Baillargeon et al., 2001; Crimmins, 1984). There have also been strident warnings about the danger this shortfall poses for North American economies in the twenty-first century, particularly given the increasing competition from the rising Asian giants, India and China. No less than the President of United States, George Bush, in his 2006 State of the Union address, took note of the problem proclaiming an "American Competitiveness Initiative"..."to give our nation's children a firm grounding in math and science" and promising to hire 70,000 new high-school mathematics and science teachers. It is to be noted that our Asian competitors have the edge not only in the number of graduates, but in the quality of those graduates. This situation must be fairly evident when even a newspaper comic strip, Doonesbury, runs a strip in which a character is described as outsourcing his own job to an Indian software engineer, but has to have the engineer mess up every few weeks so that the American employer won't guess what has happened (Trudeau, 2005).

One essential element underlying this shortage of graduates has been our failure to successfully adapt our teaching methods to the changes in students and the demands placed on them during and after their studies (Tobias, 1990; Seymour, 1992, 1995; Seymour & Hewitt, 1997). This failure to adapt how we teach is not because educational researchers have not been able to determine what kinds of changes in pedagogy would be useful (American Psychological Association, 1997). On the contrary, there is evidence that in student-centred learning environments, learners are actively-engaged and acquire improved conceptual understanding in contrast to their peers in teacher-centred learning environments (Hake, 1998a, 1998b). In a policy forum on education called "Scientific Teaching", Handelsman et al. (2004) state that "since publication of the AAAS (editor's note: American Association for the Advancement of Science) 1989 report "Science for all Americans", commissions, panels and working groups have agreed that reform in science education should be founded on "scientific teaching", in which teaching is approached with the same rigour as science at its best. Scientific teaching involves active learning strategies to engage students in the process of science and teaching methods that have been systematically tested and shown to reach diverse students." Handelsman et al. (2004) cite about a half-dozen specific examples of successful experiments at modifying teacher-centred learning environments that focus on the transmission of knowledge to something stimulating active engagement, and evidence of resultant improvements in problem-solving ability, conceptual understanding, and success in subsequent courses compared with peers experiencing traditional teacher-centred learning environments.

Despite this evidence the problem of improving science is still not solved. That is, even though some post-secondary institutions have implemented changes in the learning environment with success, such change has not become the norm in post-secondary pedagogy (Handelsman et. al., 2004). A recent study following a large cohort of students (N = 1452) through their first two years of post-secondary science studies (Rosenfield et al., 2005) assessed both faculty and students' perceptions concerning the learning environments they face in mathematics and science classrooms came to a similar conclusion. This assessment focussed primarily on the process definition of teaching. Analysis of teacher data from this study (Dedic, Dickie, Rosenfield, & Rosenfield, 2005b) showed that 36% (sample N = 84) of instructors engaged in teaching acts associated with student-centred learning environments (called in that study "fostering environments"). A sample item indicates the type of teaching acts engaged in by this group of instructors, "I encourage students to discuss ideas amongst themselves as a way to improve their understanding." while the remaining instructors did not consider such teaching acts as useful. The focus of the former group of instructors on the learning process. For example, they are significantly more likely than their peers to assess students' prior knowledge before teaching a new topic. Interestingly, the 36% of instructors who rated the environments that they created as high in "fostering" and low in "transmission" were significantly more likely to have knowledge of education research than their colleagues. These results support the claims of Handelsman et al. (2004) that required changes in learning environments by and large are not happening, and instructors are not cognizant of educational research results.

In the study by Rosenfield et al. (2005) there were two versions of a forty item assessment of learning environments instrument: one for students and one for instructors. Factor analysis of both instructor and student data revealed the same two major factors: both groups viewed the learning environments as being teacher-centred ("transmission") and/or student-centred ("fostering"). To illustrate these views, a sample item on the teacher-centred scale reads “Students should spend most of their time in class taking notes” (instructor version) or “I spent most of my time in class copying the teacher’s notes” (student version), and a sample item on the student-centred scale is “I encourage students to develop their own methods for solving typical problems” (instructor version) or “The teacher encouraged me to think for myself” (student version). A cluster analysis of student data reveals three groups of students: 1) those who perceive the environment largely as student-centred; 2) those who perceive the environment as both student-centred and teacher-centred; 3) those who perceive the environment as largely teacher-centred. Students in cluster 1) who perceived the environment as largely student-centred had significantly ($p < .001$) higher academic performance (measured by average grade in mathematics and science courses), higher affect and self-efficacy than their peers in cluster 3), and they were also more likely to persevere (Dedic, Rosenfield, Dickie, & Rosenfield, 2005a).

One further finding from the Rosenfield et al. (2005) study is that the cluster 1) students, who rated the learning environment as largely student-centred, as opposed to their peers, rated their instructors as more effective in helping them learn ($p < .001$). We reason from this that if summative assessment decisions were made on the basis of responses to the global item “The teacher was effective in making me learn.”, and teachers were made aware of the link between this global item and student-centred teaching acts, then there would be an incentive amongst instructors to move toward student-centred learning environments. This belief is fostered by the apparent ability of teachers, just like their students, to distinguish between teaching acts that are student-centred versus those that are teacher-centred.

One caveat, this research took place in four post-secondary institutions with no summative instructor evaluation policy. That is, student rating forms are not used for hiring, firing or tenure type decisions. Instead, the objective of departmental evaluation is solely to provide feedback to instructors to help guide professional development.

However, student rating forms can be an obstacle to professional development. Kolitch and Dean (1999) examined the Student Evaluation of Instruction (SEI) form used at the State University of New York, from the point of view of both the transmission or teacher-centred model of teaching and the “engaged-critical” model of teaching (their term for what was called above “fostering” or student-centred), and found implicit assumptions built into the SEI that favoured the teacher-centred model over the student-centred one. Student rating forms, which were developed to rate instruction that was presumed to use the prevailing paradigm of post-secondary instruction which is teacher-centred, that is, transmission of knowledge, have not changed over the last two decades. If non-global items from such rating forms are used for summative decisions, instructors may feel obliged to pander to the built in bias these forms exhibit towards teacher-centred learning environments, and so the rating forms themselves would become a major obstacle to adoption of more student-centred active learning strategies that educational research has shown to promote conceptual change.

Abrami, Theall and Mets (2001) raised similar concerns about the philosophies and approaches to postsecondary instruction undergoing major change from traditional didactic forms of instruction to more learner-centred approaches. When using cooperative learning techniques. We may wish to include items that assess whether the instructor facilitated positive interdependence (e.g., Were students responsible for the learning of their peers?) and individual accountability (e.g., Were teammates responsible for their individual learning?) We may also wish to ask about how well instructors facilitated problem-based inquiry and whether and how students were scaffolded to engage in self-regulated learning. In addition, the use of technology for learning, both as a way to supplement traditional instruction and as a means to deliver instruction at a distance, may require a rethinking about the qualities of effective teaching and its evaluation, especially with regard to specific teaching dimensions.

Like the perennial question about "the chicken and the egg", which comes first: changes in student rating forms or changes in post-secondary learning environments? Even as some departments or faculties move towards adding emphasis to the teaching component of tenure decisions, evaluation of teaching continues to be made on the basis of forms designed with the intention of determining if the instructor is a good transmitter of knowledge. Currently it would be foolhardy for young instructors hoping for tenure to create active-engagement learning environments that depend more on students working collaboratively, with or without technology enhancements, despite the overwhelming evidence that such active learning strategies improve learning, knowledge retention, and persistence in science studies.

These concerns reinforce our prior recommendations (Abrami et al., 1996) concerning the use of global ratings, instead of dimensional ratings, for summative purposes. In brief, we recommend that only global ratings be used for summative decisions, and that there is evidence that this can help persuade teachers to move towards more student-centred learning environments. Also, in the last ten years, with the rise of student-centered approaches, it has become clear that we need new evaluation approaches, not biased towards teacher-centred approaches, for formative assessment as well to help teachers see how to change.

USING STUDENT RATINGS FOR SUMMATIVE DECISIONS

Expert consensus regarding the use of student ratings for promotion and tenure purposes is that global ratings can reliably distinguish among outstanding, average and poor instructors. That is, broad categorizations of teaching quality are possible but fine distinctions go beyond what the instruments are capable of. Nevertheless, there is much anecdotal evidence that administrative uses of student ratings are improper, making small differences among instructors into fateful hiring and promotion recommendations. It is a problem of misplaced precision.

We are decidedly against removing human judgment from the process of judging teaching effectiveness. At the same time, we want to take the accumulated scientific evidence and see that it forms part of the decision process. Chief amongst our recommendations is to use statistical hypothesis testing to insure that judgments about excellence do not capitalize on chance fluctuation. In addition, we recommend that measurement error be more carefully reflected in how student ratings data are used.

More precisely, Abrami (2001) presented a method for insuring the appropriate precision in using student ratings for summative decisions:

1. Report the average of several global items or a weighted average of specific items, if global items are not included in the student rating form.
2. Combine the results of each faculty member's courses together. Decide in advance whether the mean will reflect the average rating for courses (i.e., unweighted mean) or the average rating for students (i.e., weighted mean).
3. Decide in advance on the policy for excluding student rating scores by choosing one of the following alternatives: a) include student ratings for all courses; b) include student ratings for

all courses after they have been taught at least once; c) include student ratings for all courses but those agreed upon in advance (e.g., exclude small seminars); or d) include student ratings for the same number of courses for all faculty (e.g., include best ten rated courses).

4. Choose between norm-referenced and criterion-referenced evaluation. If norm-referenced, select the appropriate comparison group and relative level of acceptable performance in advance. If criterion referenced, select the absolute level of acceptable performance in advance.
5. Follow the steps in statistical hypothesis testing: a) state the null hypothesis; b) state the alternative hypothesis; c) select a probability value for significance testing; d) select the appropriate statistical test; e) compute the calculated value; f) determine the critical value; g) compare the calculated and critical values in order to choose between the null and alternative hypotheses.
6. Provide descriptive and inferential statistics and illustrate them in a visual display which shows both the point estimation and interval estimation used for statistical inference.
7. Incorporate student rating validity estimates into statistical tests and confidence intervals. Norm-based statistical procedures with a correction for measurement error:

$$t_{vc} = \frac{\overline{Y}_i - \overline{Y}_g}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_g^2}{n_g}\frac{1}{1-vc}}} \text{ for df} = \mathrm{n_i} + \mathrm{n_g} - 2.$$

where $\overline{Y}$ is the mean TRF score, s^2 is the unbiased variance, n is sample size, vc is the validity coefficient, and df is the degrees of freedom.

In addition, one can calculate a confidence interval for the calculated value of t_{vc}:

$$CI = (\overline{Y}_i - \overline{Y}_g) \pm \underline{t}_{\alpha} s_{Dvc}$$

where t_{α} is critical value of t at a particular alpha level and

$$\mathrm{s_{Dvc}} = \sqrt{\frac{s_i^2}{n_i} + \frac{s_g^2}{n_g}\frac{1}{1-vc}}$$

8. Since we are interested in instructor effectiveness and not student characteristics, consider using class means and not individual students as the units of analysis.
9. Decide whether and to what extent to weigh sources of evidence other than student ratings.

If promotion and tenure committees are provided with evidence that takes into account the general impact of extraneous influences more correct decisions about teaching quality will be reached. Providing clear data and interpretative guidelines does not mean that human judgment will be ignored. Promotion and tenure committees may elect to confirm the results of statistical testing or disconfirm them, especially if a reasoned argument is provided. The wise use of statistical tools and procedures can go a long way towards overcoming the covert and overt forms of bias that characterize uneducated subjective judgment.

CONCLUSION

Ten years ago, we argued that the multidimensional nature of teaching was not uniquely captured by a single rating form and, furthermore, that dimensional ratings were highly intercorrelated. All together, we argued for the use of multidimensional ratings especially for summative decisions and for the use of global ratings.

Ten years later we have added further to that argument in two respects. First, an emphasis on student-centred learning has made traditional forms of student ratings of questionable relevance as a universal approach to judging teaching effectiveness. Second, we need to pay greater attention to how ratings are used to make summative decisions about teaching effectiveness. Global ratings are the best for doing so, especially if we provide guided and scaffolded support to their use and interpretation.

The one constancy in the twenty-first century, the Age of Information and Communication, is change, and at that, there is an increasing rate of change in how we work and what we work at. With changes in technology and work have come, perhaps unbidden, changes in societal mores and beliefs. Whatever one's moral or political philosophy, whether one is happy or unhappy with the changes that are taking place, denying either the existence of change or the large role it plays in shaping the lives and characters of our youth is pointless. Nations that fail to recognize the impact of change, and do not successfully adapt, are in danger of becoming "have nots".

REFERENCES

Abrami, P.C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P.C. Abrami and L.A. Mets (eds.), *New Directions for Institutional Research: No.109. The Student Ratings Debate: Are They Valid? How Can We Best Use Them?* (pp. 59–87). San Francisco: Jossey-Bass.

Abrami, P.C., d'Apollonia, S., and Rosenfield, S. (1996). The dimensionality of student ratings of instruction: What we know and what we do not. In J. Smart (ed.), *Higher Education: Handbook of Theory and Research* (Vol.11, pp. 213–264). New York, NY: Agathon Press.

Abrami, P.C., Theall, M., and Mets, L.A. (2001). Introduction to the student ratings debate. In M. Theall, P. Abrami and L.A. Mets (eds.), *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?* (Vol. 109, pp. 1–6). San Francisco: Jossey-Bass.

American Psychological Association. (1997). *Learner-centered Psychological Principles: A Framework for School-Redesign and Reform*. Washington, DC: http://www.apa.org/ed/lcp.html accessed July 30, 2005.

Baillargeon, G., Demers, M., Ducharme, P., Foucault, D., Lavigne, J., Lespérance, A., Lavallée, S., Ristic, B., Sylvain, G., and Vigneault, A. (2001). *Education Indicators, 2001 edition*. Québec City: Ministère de l'Éducation, Gouvernement du Québec.

Crimmins, J.C. (1984). *A Report on the Crisis in Mathematics and Science Education: What Can Be Done Now?* New York, NY: American Association for the Advancement of Science.

Dedic, H., Rosenfield, S., Dickie, L., and Rosenfield, E. (2005a). *Post-Secondary Science Students: Academic Performance, Persistence and Perceptions of the Learning Environment*, paper presented to American Educational Research Association 2006 annual meeting.

Dedic, H., Dickie, L., Rosenfield, E., and Rosenfield, S. (2005b). *Post-Secondary Science Instructors: Motivation and Perception of the Learning Environment*, paper presented to American Educational Research Association for 2006 annual meeting.

Hake, R.R. (1998a). Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics* 66(1): 64–74.

Hake, R.R. (1998b). Interactive-engagement methods in introductory mechanics courses. unpublished manuscript; on line as ref 25 at <http://www.physics.indiana.edu/~hake>, accessed Sept 21 2003.

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A, DeHaan, R., Gentile, J., Lauffer, S., Stewart, J., Tilghman, S., and Wood, W. (2004). Scientific Teaching, a policy forum in *Science* 304: 521–522.

Kolitch, E., and Dean, A.V. (1999). Student ratings of instruction in the USA: Hidden assumptions and missing conceptions about "good" teaching. *Studies in Higher Education*. Routledge, Vol. 24, #1, pp. 27–42.

OECD. (2005). Education at a Glance 2005. 520 pp., ISBN 9264011919 http://www.oecd.org/document/34/0,2340,en_2649_34515_35289570_1_1_1_1,00.html accessed January 31, 2006.

Rosenfield, S., Dedic, H., Dickie, L., Rosenfield, E., Aulls, M.W., Koestner, R., Krishtalka, A., Milkman, K., and Abrami, P. (2005). *Étude des facteurs aptes à influencer la réussite et la rétention dans les programmes de la science aux cégeps* *anglophones*, Final Report submitted to Fonds de recherche sur la société et la culture, October 2005, at <http://sun4.vaniercollege.qc.ca/fqrsc/reports/fr_22.pdf>, accessed October 31, 2005.

Seymour, E. (1992). "The Problem Iceberg" in science, mathematics, and engineering education: Student explanations for high attrition rates. *Journal of College Science Teaching* 21: 230–238.

Seymour, E. (1995). Revisiting the 'Problem Iceberg': Science, Mathematics, and Engineering Students Still Chilled Out, *Journal of College Science Teaching* 24(6): 392.

Seymour, E. and Hewitt, N. (1997). *Talking about Leaving: Why Undergraduates Leave the Sciences*. Boulder, CO: Westview.

Tobias, S. (1990). *They're not Dumb, They're Different: Stalking the Second Tier. An Occasional Paper on Neglected Problems in Science Education*. Tucson, AR: Research Corporation.

Trudeau, G.B. (2005) Doonesbury, Flashback, *The Montreal Gazette*, October 29.