Journal of Educational Psychology

Dimensionality, Reliability, Validity, Potential Biases, and Utility

Herbert W. Marsh

of teacher and student) context variables different approaches This process serves

jngs and to explore directions for future re- evaluation items. Poorly worded or inap-

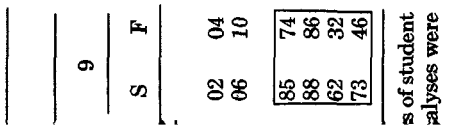
the construct validation approach described formation. Student ratings, like the teach-

uses in the research and development of the criteria being considered (Marsh &

[redacted]

Table 1
Factor Analyses of Students' Evaluations of Teaching Effectiveness (S) and the Corresponding Faculty Self-Evaluations of Their Own Teaching (F) in 329 Courses

Evaluation items (paraphrased)	Factor pattern loadings																	
	1		2		3		4		5		6		7		8		9	
	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F
1. Learning/Value	[redacted]																	



alizable (e.g., a teacher who was judged to be well organized but lacking enthusiasm in one course was likely to receive a similar pattern of ratings in other classes). These findings

Table 2

[redacted]

large impact on factor analyses of individual each provided clear support for the multi-

from 25 students, .74 from 10 students, .60 from five students, and only .23 for one stu-

Marsh & Overall, 1979a) demonstrated that consistent with previous research, the sin-

Table 3
Correlations Among Different Sets of Classes for Student Ratings and Background Characteristics

Measure	Same teacher, same course	Same teacher, different course	Different teacher, same course	Different teacher, different courses
Enthusiasm	.734	.613	.011	.028
Organization/Clarity	.676	.540	−.023	−.063
Group Interaction	.699	.540	.291	.224
Individual Rapport	.726	.542	.180	.146
Breadth of Coverage	.727	.481	.117	.067
Examinations/Grading	.633	.512	.066	−.004
Assignments	.681	.428	.332	.112
Workload/Difficulty	.733	.400	.392	.215
Overall course	.712	.591	−.011	−.065
Overall instructor	.719	.607	−.051	−.059
Mean coefficient	.707	.523	.140	.061
Background characteristic				
Reason for taking course (percent indicating general interest)	.770	.448	.671	.383
Class average expected grade	.709	.405	.483	.356
Workload/difficulty	.773	.400	.392	.215
Course enrollment	.846	.312	.593	.058
Percent attendance on day evaluations administered	.406	.164	.214	.045
Mean coefficient	.690	.340	.491	.211

lidity. The most widely accepted criterion of effective teaching is student learning, but other criteria include changes in student behaviors, instructor self-evaluations, the evaluations of peers and/or administrators who actually attend class sessions, the frequency of occurrence of specific behaviors

First, the ratings were not of the instructor in charge of the course but of teaching assistants who played a small ancillary role in the actual instruction. Thus, there was no way to separate achievement produced by a teaching assistant from that due to the instructor; a student who put too much reli-

observed by trained observers, and the effects of experimental manipulations.

Multisection Validity Studies

of lectures by the instructor might evaluate the assistant highly and perform poorly on the exam. Doyle (1975) also argued that a

[redacted]

lidity studies is more adequate, numerous [redacted] it is important to have effective pretest

identified an alternative explanation for ... tion average expected grades indicated by

of former students that are unlikely to con- faculty self-evaluations in some areas but

Table 4
Multitrait-Multimethod Matrix: Correlations Between Student and Faculty Self-Evaluations in 329 Courses

	Instructor self-evaluation factor									Student evaluation factor								
Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Instructor self-evaluations																		
1. Learning/Value	(83)																	
2. Enthusiasm	29	(82)																
3. Organization	12	01	(74)															
4. Group Interaction	01	03	−15	(90)														
5. Individual Rapport	−07	−01	07	02	(82)													
6. Breadth	13	12	13	11	−01	(84)												
7. Examinations	−01	08	26	09	15	20	(76)											
8. [illegible]	[illegible]																	

(1975) compared peer ratings based on classroom visitation and student ratings at a brand new university, thus reducing the probable confounding of the two sources of information. Three different peers evalu-

less sensitive, reliable, and valid; (2) more threatening and disruptive of faculty morale; and (3) more affected by non-instructional factors such as research productivity" (p. 45) than student ratings.

was a relative lack of agreement among peers

Behavioral Observations by External

their value as a criterion of effective teaching.

significantly differentiated among the three

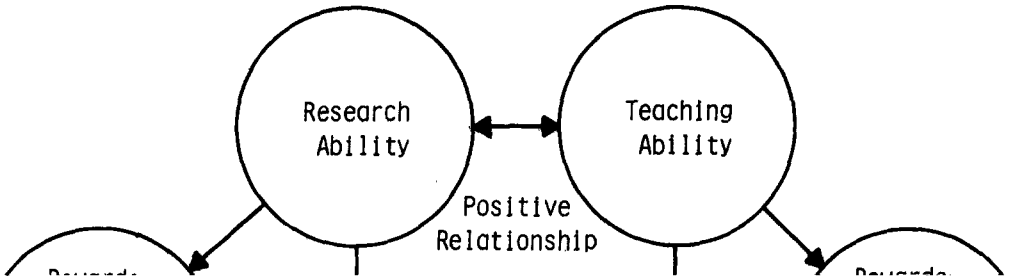
(in press), and Rosenshine and Furst (1973)

achievement. Both naturalistic observations and experimental manipulations of clarity-related behaviors are significantly

ratings, specific behaviors and observational factors must also be related to external indicators of effective teaching. In this re-

correlated with student ratings and with

spect, the research conducted on teacher


Research
Ability
Teaching
Ability
Positive
Relationship

validity, are so willing to accept other indicators that have not been tested or have been shown to have little validity.

Relation to Background Characteristics:

course (e.g., class size, content area, students' interest in the subject, etc.) and to rate the "ease of teaching this particular course." These ratings of ease-of-teaching (see Table 6) were not significantly correlated with any of the student rating factors and were nearly

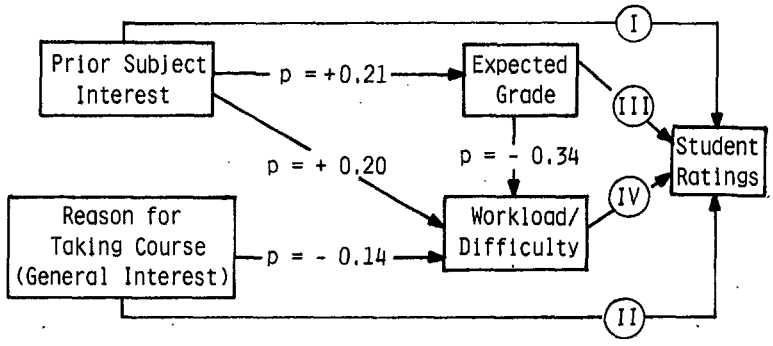


Figure 2. Path analysis model relating prior subject interest, reason for taking course, expected grade, and Workload/Difficulty. (Path coefficients for the student rating factors appear in Table 5.)

similar background characteristics of hours ... general interest only. A path analysis (see

Table 5
Path Analysis Model Relating Prior Subject Interest, Reason for Taking Course, Expected

Grade and Workload/Difficulty to Student Ratings

	Factor											
	Interest			Interest Only)			Course Grade			Difficulty		
Student ratings	DC	TC	Orig r	DC	TC	Orig r	DC	TC	Orig r	DC	TC	Orig r
Learning/Value	36	44	44	15	13	15	26	20	29	17	17	12

from correlations obtained when the analysis is performed on responses by individual students. Hence, effects based on individual

constitutes a bias. Alternative definitions of bias, which are generally implicit rather than explicit, are described below.

ence. For example, even though student learning in multisection validity studies is defining bias by statistically controlling for potential biases with multiple regression

correlated with student ratings, this effect should not be considered a bias. However, techniques or by forming normative (cohort) groups that are homogeneous with respect

[redacted]

of correlations between a specific variable and the set of student evaluation factors is

was moderately correlated with Group In-

[redacted]

very large classes can free up enormous amounts of instructional time that can be used to substantially reduce the average class size in the range where the effect of class size

the student ratings and instructor self-evaluations. Higher student interest in the subject apparently creates a more favorable learning environment and facilitates effec-

does appear to be negative. However, I (Marsh, Overall, & Kesler, 1979a) argued that my correlational effect should be interpreted cautiously and speculated that the unexpectedly higher ratings for very large classes could be due to (a) the selection of particularly effective instructors with dem-

tive teaching, and this effect is reflected in student ratings as well as faculty self-evaluations.

Workload/Difficulty. The Workload/Difficulty effect on students' evaluations was also one of the largest found (Marsh, 1980b, 1983). Paradoxically, at least based on the

onstrated success in such settings; (b) students systematically selecting classes taught

supposition that Workload/Difficulty is a potential bias to student ratings, higher

Table 6

[redacted]

0.000 Tc(e) Ti 0.124 Tw106.000 Tz 0.184 Tc(Ratin) Ti0.000 Tc(e) Ti 310 Tc(ET8((actua) Ti0.000RTc(l) Ti-0.583 Tw99.557180.178 LearTr0.000 0.000 n.184 Tc(Ratin) Ti0.000TT0.927i3 Tr10.000 T

[redacted]

Marsh, Fleiner, and Thomas (1975) and Marsh and Overall (1980) examined class-average pretest scores, expected grades, each class who received grades and those who did not, and there was substantial agreement with evaluations by the two

multisection validity studies described ear- class-average grades of those students who

below) in which grading standards were experimentally manipulated. Groups of students viewed a videotaped lecture, rated teacher effectiveness, and took an objective

ratings, support for this suggestion is weak and the size of such an effect is likely to be insubstantial in the actual use of student ratings.

were given their examination results and a

Table 7
Overview of Relations Found Between Students' Evaluations of Teaching Effectiveness and Specific Background Characteristics

Background Characteristic	Summary of "Typical" Findings
	it is not always clear if interest existed before the start of course or was generated by the instructor.
Expected/actual grades	Classes expecting (or actually receiving) higher grades give somewhat higher ratings, though this can be interpreted to mean either that higher

and teaching effectiveness was evaluated. Despite the fact that the lecture content was in the way they were affected by the experimental manipulations. In the condition

specifically designed to have little educa- most like the university classroom in which

tional value, the ratings were favorable. The students were told before viewing the lecture

Consistent with the Marsh and Ware reanalysis, they also found that in the few studies that analyzed separate rating factors, the rating factors that were most logically related to the expressiveness manipulation were most affected by it. Finally, they

tifaceted ratings in this article, a particularly powerful test of the validity of student ratings would be to show that each rating factor is strongly influenced by manipulations most logically associated with it and less influenced by other manipulations. This is

manipulation did interact with the content manipulation and a host of other variables

reanalysis of the Dr. Fox data described above, and it offers strong support for the

[redacted]

more accurately predicted by differentially [redacted] university classes the frame of reference is [redacted]

taped lectures seems dubious). Unfortunately, the effects of content and expression-Rose & Menges, 1981). SEEQ has been used in two such studies using multiple sec-

the effect of consultation without feedback other indicators of teaching effectiveness in

(i.e., a placebo effect due to consultation, or evaluating total faculty performance in

a real effect due to consultation that does not depend on feedback from student ratings). Second, the criterion of effective teaching used to evaluate the studies was limited

North American universities (for reviews see Centra, 1979; Leventhal et al., 1981; Seldin, 1975). Each survey found that classroom teaching was considered to be the most im-

primarily to student ratings; only the Overall important criterion of total effectiveness

able for subjects in these studies to assume that the chairperson's report was at least partially based on students' evaluations.

These studies demonstrate the importance of reports of teaching effectiveness but do

marized by a single score representing an optimally weighted average of specific components or by the separate presentation of each of the multiple components, but there is no research to indicate which is most

Overview, Summary, and Implications

[redacted]

have a systematic voice in the interpretation of their student ratings.) Consequently, [redacted]

[redacted]

onstrates that student ratings are clearly [redacted] cators of effective teaching are proposed [redacted]

multidimensional, quite reliable, reasonably valid, relatively uncontaminated by many

(e.g., Centra, 1979), few are supported by systematic research, and none are as clearly

considerable base of research from which to form opinions about their worth. However

on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand

lege faculty. *Journal of Higher Education, 46,* 89–102.

Marsh, H. W. (1977). The validity of students' evaluations: [illegible]

man's "Consistency and variability among college students in rating their teachers and courses." *Research in Higher Education, 10,* 139–147.

lecturer expressiveness, and density of lecture content on student ratings. *Journal of Educational Psy-*

81–91.

Price, J. R., & Magoon, A. J. (1971). Predictors of

Menges, R. J. (1973). The new reporters: Students rate instruction. In C. R. Pace (Ed.), *Evaluating learning and teaching.* San Francisco: Jossey-Bass.

Proceedings of the 79th annual convention of the American Psychological Association, 7, 523–524.

Remmers, H. H. (1963). Teaching methods in research on teaching. In N. L. Gage (Ed.), *Handbook*

Student achievement as a measure of instructional

Rodin, M., & Rodin, B. (1972). Student evaluations of teachers. *Science, 177*, 1164–1166.

47, 79–88.

Rosenshine, B. (1971). *Teaching behaviors and student achievement.* London: National Foundation

Warrington, W. G. (1973). Student evaluation of instruction at Michigan State University. In A. J. [illegible] student ratings of instruction under different incentive conditions: A further study of the Dr. Fox