Journal of Marketing Education
Volume 31 Number 1
April 2009 16-30

10.1177/0273475308324086
http://jmd.sagepub.com
hosted at
http://online.sagepub.com

Student Evaluations of Teaching: Are They Related to What Students Learn?

A Meta-Analysis and Review of the Literature

Dennis E. Clayson
University of Northern Iowa, Cedar Falls

Although the student evaluation of teaching has been extensively researched, no general consensus has been reached about the validity of the process. One contentious issue has been the relationship between the evaluations and learning. If good instruction increases the amount of learning that takes place, then learning and the evaluations should be validly related to each other. A review of the literature shows that attempts to find such a nomological relationship has been complicated by practice, methodology, and interpretation. A meta-analysis of the literature shows that a small average relationship exists between learning and the evaluations but that the association is situational and not applicable to all teachers, academic disciplines, or levels of instruction. It is concluded that the more objectively learning is measured, the less likely it is to be related to the evaluations.

Keywords: *student evaluation of teaching; learning; grades; rigor; business education*

Few issues within academics have been as well researched, documented, and long lasting as the debate about the student evaluation of teaching (SET). The first published article on the evaluations was written by researchers from Purdue University more than 80 years ago (Remmers & Brandenburg, 1927, as cited in Kulik, 2001). In the early 1950s, the University of Washington became one of the first institutions to begin conducting a formal evaluation analysis (Guthrie, 1954). Since the 1970s, which saw almost an explosion of research, the application of SET has become nearly universal. By the 1990s, four of five campuses used some sort of SET (Seldin, 1993). Perhaps because of the appreciation of feedback in business and in research on consumers, business schools have been particularly heavy users of the evaluations. Currently, almost all business schools (99.3%) use some form of SET, and deans generally place a higher importance on these than either administrative or peer evaluations (Comm & Manthaisel, 1998). On many campuses, student evaluations of instruction are the most important and, in many cases, the only measure of teaching ability (Wilson, 1998). Seldin (1999) reported a California dean as saying, "If I trust one source of data on teaching performance, I trust the students" (p. 15).

As would be expected of a process that can establish tenure, promotion, merit pay, and reputation, SET has been widely debated and researched. Nevertheless, little agreement has been made on key points. This article reviews the issue from the one area that has the most potential for explaining many of the differences that have been found in the literature and that would have a direct impact on the actual practice of the evaluation of business education. Simply put, are the evaluations that students make of courses and instructors related to student learning? This study has three parts. First, a review of the history and issues involved in the learning/evaluation debate is presented. Second, the extant literature was utilized to create a meta-analysis of articles specific to the learning/evaluations question. And third, a conclusion was drawn from the literature review and meta-analysis of the nature of the relationship, and the implications of that conclusion are given for marketing education and further research.

Background and Issues

Much of the debate about SET has been not about the need for evaluation, but about whether the instruments have a valid application for this purpose. The related line of research has been handicapped by a fundamental problem. Essentially, no one has given a widely accepted definition of what "good" teaching is, nor has a universally agreeable criterion of teaching effectiveness been established (J. V. Adams, 1997; Kulik, 2001).

Author's Note: Please address correspondence to Dennis E. Clayson, Professor of Marketing, University of Northern Iowa, Cedar Falls, IA 50614-0126; e-mail: dennis.clayson@uni.edu.

Nevertheless, both defenders and detractors of SET generally agree that students will learn more from good teachers. In other words, if the process is valid, then there should be an association between student learning and the evaluations that students give of classes and instructors. As with any topic in education, there are those who disagree. Scriven (1983) observed, "The best teaching is not that which produces the most learning" (p. 248), but this is a minority position. Cohen (1981) affirms, "Even though there is a lack of unanimity on a definition of good teaching, most researchers in this area agree that student learning is the most important criterion of teaching effectiveness" (p. 283).

When looking at the literature, there are several issues that become important when attempting to evaluate the learning/SET association.

Situational Problems

As indicated earlier, research findings have been variable and controversial. Some background of the nature of the resultant controversy is necessary to fully understand the nature of the literature base. The defenders of the SET process are generally found in the colleges of education, in the national teachers' unions, and among those who consult in the area. Their positive attitude toward SET is compatible with a holistic environment that consists of positive research findings, currently accepted educational philosophy, and a communication system largely centered within their own academic disciplines. They are confident enough in their positive conclusions to dismiss negative findings as "myths" (Aleamoni, 1999; Marsh & Roche, 2000) and to wonder why negative comments continue to be found in the literature (Theall & Franklin, 2001). Because instructional-related research is the province of their occupation, it would be expected that the majority of the research on SET would come from those in educational disciplines.

Academic areas outside of education tend to look on research in education and instruction as less prestigious. Noneducational researchers' excursions into this area are typically seen as an interruption from their primary research interests and many times is motivated by a pragmatic response to specific pedagogical problems or concerns. These researchers are more scattered and isolated throughout the academic disciplines and have a much narrower range of publication outlets. Although there are a limited number of excellent journals specializing in business education—such as *Research in Economic Education*, *Academy of Management Learning & Education*, and the *Journal of Marketing Education*—research published in these sources are seldom cited in the leading journals of the education disciplines.

As an example, the references from four sources can be compared. Cashin (1988, 1995) prepared a series of short reviews of SET research that are typical of the summaries offered to instructors in education. His 1995 summary consists of 67 references, only 1 of which came from a researcher outside of the educational disciplines. Sixty-four percent of the first authors are cited repeatedly. One frequently cited article published by Marsh and Roche (2000) in the *Journal of Educational Psychology* contains 59 references, all from education, statistics, or psychology. Sixty-one percent of the first authors are cited more than once. The first author cites himself 17 times, which is a rough measure of the researcher's prior interest and publication success in the area. Compare these to a well-received research article on SET by Marks (2000) published in the *Journal of Marketing Education*. His article has 79 references; 49% are from educational journals, and 16% are from business journals. Thirty-three percent of the authors are cited repeatedly, and the author never cites himself. A look at the *Journal of Marketing Education* over the last 5 years shows nine articles on or related to SET (Clayson, 2007; Clayson & Sheffet 2006; Faranda & Clarke, 2004; Gremler & McCollough, 2002; Laverie, 2002; Paswan &Young, 2002; Schlee, 2005; Schmidt, Houston, Bettencourt, & Boughton, 2003; Wilhelm, 2004). Of these 15 authors, only 4 cited any previously published research related to their current research on SET, and all were published in business education journals.

The researchers outside of the education disciplines are, however, equally emphatic and direct in their views toward SET. Sheets, Topping, and Hoftyzer (1995) replied to education researchers who claim that students are able to judge the quality of their instruction with the comment "The basis for this view amounts to little more than the belief that correlation proves causation" (p. 55). In such an environment, with so many contradictory findings (Dowell & Neal, 1982), it becomes relatively easy to select research that reinforces a point of view. It is a mistake to look at one or two articles about SET in one's own discipline when looking for direction on the evaluations. On the other hand, it is the author's opinion that a skewed view will result from looking only at educational summaries for direction, without critical evaluation of original sources, and without looking for research from one's own academic discipline.

Methodological Problems

Irrespective of disciplinary differences, research must begin somewhere. This was recognized by Cohen (1981), who explains:

> It [teaching effectiveness] can be further operationalized as the amount students learn in a particular course. This is at best a crude index of teaching effectiveness because a number of factors outside the teacher's control—student ability and motivation, for example—affect the amount students learn. Nonetheless, if student ratings are to have any utility in evaluating teaching, they must show at least a moderately strong relationship to this index. (p. 281)

If both learning and SET are related to good teaching, then SET should be found to be related to learning. A test of this assertion has been hindered by several methodological difficulties, the most fundamental of which is how learning can be measured. Most of the issues can be summarized within three areas: the grade/SET relationship, student perception of learning, and the relationship of actual grades to learning.

The grade/SET relationship. Grades are valuable to students. They are willing to work for them. Students are not always willing to work to learn more. The extent that students give higher evaluations to instructors who give them a higher grade may or may not be an indication of a learning/SET association. The extent to which higher evaluations can be "bought" by grades, which are not directly related to learning, actually invalidates SET. This issue has had such historical importance that it needs to be briefly dealt with before a discussion of the relationship between learning and SET can be adequately addressed.

SET defenders admit there is an apparent grade/evaluation association (Braskamp & Ory, 1994; Cashin, 1995; Marsh & Dunkin, 1992), but many continue to assert that grading standards do not significantly change SET, especially if other variables such as rigor and prior student interest are taken into account and properly controlled (Cashin, 1995; Greenwald & Gillmore, 1997a; Kaplan, Mets, & Cook, 2000; Marsh & Roche, 2000; Powell, 1977; Schwab, 1976; Seiver, 1983; Stumpf & Freedman, 1979). Some simply claim that the grade/evaluation correlation is too small to be meaningful (Marsh & Roche, 1999) or that grades and good teaching are validly related (Cohen, 1981; Marsh, Hau, Chung, & Siu, 1997; Marsh & Roche, 1997).

Many detractors maintain that grades influence SET in a fashion that invalidates the instruments (Gillmore & Greenwald, 1999; Greenwald & Gillmore, 1997b; also see Greenwald, 1997, for use of data). In business classes, expected grades have been found to create a significant difference in the evaluations of instructors (Bharadwaj, Futrell, & Kantak, 1993; Goldberg & Callahan, 1991). Two negative hypotheses have been advanced. The oldest of the two (*leniency*) states that students give lenient-grading instructors higher evaluations. A number of researchers have found evidence of such an effect (Greenwald & Gillmore, 1997b; see Johnson, 2003, for an extensive review). A second hypothesis (*reciprocity*) states that students who receive better grades give better evaluations, irrespective of any leniency tendency of the instructor (Clayson, 2004; Clayson, Frost, & Sheffet, 2005). The two are distinctly different hypothetical and statistical concepts, but they are not necessarily mutually exclusive (Clayson, 2007; Stumpf & Freedman, 1979). Johnson (2003) found strong evidence for both a leniency and a reciprocity effect. Another large study found a robust leniency effect in accounting classes even when controlling for other measures of learning (Weinberg, Fleisher, & Hashimoto, 2007). Marsh and Roche (2000) and Centra (2003) also performed large careful studies but did not find evidence for the leniency effect. The differences in findings may be the result of problems generated by utilizing only between-class data, which is both statistical and logical (Clayson, 2007). If class means are utilized, then reciprocity and leniency effects become confounded. Leniency effects can be found that do not exist at the student level, unless class size and average grades are held constant. Irrespective of the demands of holistic models, instructors are ultimately evaluated by students, not by classes. Clayson et al. (2005) maintain that the only hypothesis supported by all the data is reciprocity.

Ironically, grades may negatively affect SET irrespective of which side of the issue is most correct. Although the actual relationship can be debated, it does not change the behavior of both faculty and students who believe that there is a relationship between grades and the evaluation. There is evidence that the belief alone modifies faculty and student behavior (Birnbaum, 2000; Goldman, 1985; Kolevzon, 1981; Marsh, 1987; Moore & Trahan, 1998; Redding, 1998; Ryan, Anderson, & Birchler, 1980; Simpson & Siguaw, 2000).

Both sides of the debate apparently agree that grades, as utilized in these studies, are either only marginally related to learning or not related at all. Nor do these researchers claim that students need to believe that their grade is equivalent to their perception of learning. Cashin (1995) sums up the thinking of many when he maintains that grades cannot be used as an indicator of learning without proper control. In this, he is in agreement with researchers outside of educational disciplines (Clayson, 2004; Greenwald & Gillmore, 1997a; Johnson, 2003).

Student perception of learning. The second methodological problem arises from utilizing student perceptions as a measure of learning. Some researchers have asserted that students are the best judge of what they are learning (Cruse, 1987; Machina, 1987). This claim has not been widely supported by actual findings. Students' perceived grades need not be strongly related to their actual grades (Baird, 1987; Clayson, 2005a; Flowers, Osterlind, Pascarella, & Pierson, 2001; Sheehan & DuPrey, 1999; Williams & Ceci, 1997). Students make consistent errors in estimating their grades that have been interpreted as a "metacognitive" effect. In other words, poorer students don't know what they don't know and consequently overestimate their knowledge of tested material, whereas better students know what they don't know and underestimate their knowledge (Grimes, 2002; Kennedy, Lawton, & Plumlee, 2002; Moreland, Miller, & Laucka, 1981). A more recent finding indicates that this interpretation may be oversimplified. Students appear to know what they don't know, but they utilize this recognition only as an averaging foundation that is combined with group norms to estimate their learning performance (Clayson, 2005a).

Actual grades and learning. The use of actual course grades also creates problems. Intensified by grade inflation, there are numerous variables that can affect course grades unrelated to student skills and knowledge that without strong statistical or experimental controls make the grades questionable as a measure of learning. Pollio and Beck (2000) point out that "The notion that grades provide accurate indices of how well a student is doing in college and how well he or she will do in a future career is not supported by the empirical literature" (p. 100). An anonymous reviewer from a leading education journal stated to the writer, "I do not think that the phrase 'learning' can be proxied [*sic*] by the course grade."

Methodological Solutions

At least five solutions have been advanced in response to Cashin's (1995) call for stringent controls when measuring learning.

1. It has been suggested that the grade variable utilized to measure learning should be from class means and not from individual students (Abrami, d'Appolonia, & Cohen, 1990; Cohen, 1981; Marsh & Roche, 2000). This removes, or averages, individual differences and reflects most closely the instructional impact of the class and/or instructor. As mentioned earlier, this solution has been criticized because it can result in interactions that cannot be easily untangled statistically. The variability between students, even though averaged, can confound the variability between group means (Weinberg et al., 2007). It is possible for individual students not to show any relationship between the evaluations and learning, yet the between-class mean data could show a significant relationship (Clayson, 2007).
2. Common tests could be used in multiple sections, especially if the variance between instructors can be controlled (Cohen, 1981; Williams & Ceci, 1997).
3. The measure of learning could be the change in grades based on pretest and posttest conditions (Hake, 2002).
4. Learning could be measured by the performance in future classes controlled for student characteristics and performance in prerequisite classes (Johnson, 2003; Weinberg et al., 2007; Yunker & Yunker, 2003).
5. Surprisingly, from a methodological viewpoint, few sources were found to recommend or utilize outside standardized measures. This is probably due to practical considerations. Standardized learning measures are relatively rare and, to be useful, standardized tests would need to be appropriate both for the subject matter and the academic level of any given course. The one notable exception is the use of the *Test of Understanding in College Economics* (TUCE) in economic SET studies (Marlin & Niss, 1980; Soper, 1973).

The Paradox of Rigor

The interpretations of the grading/SET studies are complicated by another finding. Most of the literature has found an apparent inconsistent relationship between rigor and the evaluations. If effort and challenge are related to learning, and learning is related to good teaching, then as reasonable levels of effort and challenge (rigor) increase, the overall level of the evaluation (SET) should increase as well (Greenwald & Gillmore, 1997a). Those who summarize SET research within educational disciplines generally agree (Cashin, 1995; see Sixbury & Cashin, 1995). Measures of rigor (workload, difficulty of material, time, etc.) are not included in all studies of the learning/SET relationship, but when they are added, the association between rigor and SET is generally negative (Attiyeh & Lumsden, 1972; Clayson & Haley, 1990; Frey, Leonard, & Beatty, 1975; Steiner, Holley, Gerdes, & Cambell, 2006). Centra (2003) found a negative association between "student effort/involvement" and the evaluations. In Cohen's (1981) meta-analysis, "difficulty" was found to be unrelated to measures of learning. In 24 studies, Cohen found only one significant relationship between difficulty and learning, and that was negative. Weinberg et al. (2007) put the relationship into economic terms when they stated, "greater human capital production is associated with less pleasant course experience for students because more work is required of them" (p. 7). Two caveats are warranted in this discussion. In some studies, the operational definition of *difficulty* could be seen as negative, indicating that the instructor was in violation of some norm. Second, a student would be logically expected to learn poorly if the instruction were not rigorous enough or if it became too rigorous.

Johnson (2003) overcame these problems by stepping around the operational definition issue and looking at the actual results of his construct. He reported that "stringent grading is associated with higher levels of achievement in follow-on courses" (p. 161), but stringent grading was strongly associated with lower evaluations. He also found that items that had a high association with the evaluations, such as instructor concern, hours per week spent in class, knowledge of course goals, and effectiveness of exams, were not related to measures of student learning. Chacko (1983), utilizing an experimental design, found that a treatment group that was graded more stringently gave significantly lower evaluations of the teacher on preparation, knowledge of subject matter, and intellectual motivation than the control group. The instructor was also rated lower in personal characteristics such as self-reliance, confidence, and sense of humor.

In summary, both defenders and detractors have consistently found positive relationships between the students'

perceptions of their own learning with their evaluations of instruction, while finding negative associations between their perception (including related behavior) of rigor and SET.

Explanation of the Rigor Paradox

There are at least five possible explanations for this phenomenon. First, the results may be a methodological artifact. Research in SET has been criticized for not utilizing enough statistical control (Clayson, 1994; Gaski, 1987; Howard & Maxwell, 1980; Seiver, 1983). Rigor's relationship to the evaluation may also depend on when it is measured. A hint is found in studies conducted by the writer (unpublished). When rigor was measured during the course of the term, the relationship was negatively related to the evaluation. When the measurement was made after the students had completed the course, the association became positive.

Second, rigor-related data has both linear and curvilinear components, suggesting that samples could show different associations depending on the level of rigor (Marsh & Roche, 2000). Students appear to give the highest evaluations to rigor that they perceive as being appropriate (Paswan & Young, 2002). As Centra (2003) suggests, "What these findings indicate is that teachers will receive better evaluations when their courses are manageable for students" (p. 515).

Third, students can also show their reaction to rigor by voting with their feet, which could skew rigor's statistical effects in any given sample. With a wide selection of majors, minors, and even within-major class sections, students may self-select classes based on their own preferences for rigor. Wilhelm (2004) compared course evaluations, course worth, grading leniency, and course workload as factors of business students choosing classes. Her findings indicated that "students are 10 times more likely to choose a course with a lenient grader, all else being equal" (p. 24). Johnson (2003) found the influence of grading policies on student course selection decisions to be substantial, even for average grade differences as small as between B+ and B.

Fourth, the perception of rigor and learning may be recursive. If students believe that they have learned well, then rigor would be perceived as being at an appropriate level. If perceived learning was low, then rigor, at whatever level, may be perceived as being inappropriate. Students' perceptions of grades are also related to their perception of the personality of the instructor (Clayson & Sheffet, 2006). Consequently, if this suggestion is correct, we would expect a grade/rigor/personality/evaluation association. This is exactly what has been found with marketing students. Clayson and Haley (1990) found rigor, even defined in rather negative terms, to be significantly and positively related to the students' perceptions of learning, but negatively linked to instructional fairness, which made its total effect on the evaluation negative. Marks (2000) replicated this study with "students enrolled in business courses" and found similar results. Bacon and Novotny (2002) noted that rigor interacts not only with the students' perceptions of the instructor, but also with the students' own personalities. A lenient instructor would increase evaluations by attracting low-achievement-striving students, but less so with highly motivated students.

There is a fifth explanation that is more holistic or philosophical than the ones above. How is education viewed? It is not necessarily true that students' expectations of the relationships found important by researchers are compatible with the underlying assumptions of these researchers. A sample of psychology students found that 65% wanted "success" as an outcome of a class, but only 35% defined that as learning (Gaultney & Cann, 2001). In another study, less than 2% of students believed a class member should fail a liberal arts class if he or she failed to perform satisfactorily and failed to meet minimum class requirements, as long as he or she put in effort in the class (J. B. Adams, 2005). A survey of 750 freshmen in business classes revealed that almost 86% did not equate educational excellence with learning. More than 96% of the students did not cite "knowledgeable" as a desirable quality of a good instructor (Chonko, Tanner, & Davis, 2002). Students do not generally believe that a demand for rigor is an important characteristic of a good teacher (Boex, 2000; Chonko et al., 2002; Clayson, 2005b). Furthermore, students seem to have decoupled their perception of grades from study habits. In management classes, no correlation was found between study production and learning production, "meaning students did not necessarily think they learned more in courses in which they studied more" (Stapleton & Murkison, 2001, p. 281). Johnson (2003) writes, "In any case, the lack of an association between grades and study habits suggests that the motivating effect of grades on student effort is not fully understood" (p. 79).

Meta-Analysis: Learning/SET Association

Because a number of articles relating learning to SET exist, and many of the findings are contradictory, a meta-analysis was conducted to investigate published findings of a learning/SET association.

Method

The meta-analysis was conducted consistent with procedures recommended by investigators who have studied the problem of combining data from numerous published results (Hunter & Schmidt, 2004; Lyons, 1997: Rosenthal, Rosnow, & Rubin, 2000; Schulze, 2007).

Locating studies. An inspection of the educational and noneducational literature related to SET was conducted. First, all the references from established articles were inspected, and reference branching was conducted. Second, all current (within the last 5 years) issues of business educational journals were inspected. Third, major databases such as Education Full Text, ERIC (EBSCOhost), PsycARTICLES, PsycINFO, ABI/INFORM, Business Source Elite, SpringerLink, and PubMed were inspected (a full list of reference sources utilized can be found at http://www.library.uni.edu/gateway/ml/find.php). Articles specific to learning and SET were selected from this large sample of studies. This is the first meta-analysis to combine business and education sources together.

Criteria for including studies. Generally, the same criteria as in the historical meta-analysis by Cohen (1981) were utilized, with several additions. First, the unit of analysis had to be directly related to college instruction. For example, an early study used by some meta-analyses was eliminated because the classes were conducted at a military base and the course lasted only 8 days (Morsh, Burgess, & Smith, 1956). Second, the data had to be based on multiple sections of the same class. Third, the measure of learning had to be common across all sections. Fourth, the learning measure had to be based not on the perception of the students, but on actual testing results. The otherwise acceptable research of Baird (1987) and Steiner et al. (2006) was excluded because they utilized "How much did you learn . . ." "about this subject" (p. 91) or "in this class" (p. 361), respectively, as the measure of learning. Fifth, the evaluation had to be conducted before the student completed the common learning instrument. Several articles were excluded for other reasons. For example, one meta-analysis (Abrami, Cohen, & d'Apollonia, 1988) was not utilized because all the sources they reviewed were already included in the present study.

In all, 17 articles were found that contained 42 studies, including 1,115 sections. The median number of sections per study was 14. The first article was published in 1953, and the last appeared in 2007. Although recent studies have been found, the majority of the research was conducted in the 1970s. The studies, their references, and their characteristics are listed in Tables 1 and 3. In addition, 7 articles were found reporting 11 studies of students summed across sections (within-class data); they are shown in Table 6.

Identification of possible mediating variables. An inspection of Table 1 indicates a wide range of associations that may be accounted for by other intermediating factors in the learning/SET association. For example, note the study by Sheets et al. (1995); one study of 58 sections of microeconomics had an average correlation of .177. Yet in the 63 sections of macroeconomics reported in the same study, the average correlation was −.142. Sheets et al. point out that the two classes are dissimilar. The second was composed mostly of freshmen and satisfied general education credit. The first was taken by more advanced students satisfying a business major requirement.

Given the discussion at the beginning of this report, whether a study came from an educational and/or psychological discipline was noted. In Table 3, the studies were also coded by the type of statistical control utilized in measuring the association, the academic discipline of the class, and the objectivity of common examinations measured on a verbal to an applied quantitative continuum (*objectivity*). The last is justified by the following: The content of the common exams was typically not given, so three general categories were created: (1) The class was in a discipline that generally required answers on a common exam that could be verbally expressed, (2) the class required math to find an objective answer, and (3) the class utilized math as a method of understanding and manipulating other learned material, such as physics or accounting. Although the relationship is not perfect, it was thought that this breakdown would roughly reflect the objectivity of the skills required to succeed. The exams in the first type of classes were more likely to utilize answers that could be recognized (multiple choice) and memorized, and that may be more subjectively evaluated. The second type of class would be more likely to have exams that require a calculation to find an answer, and the answer is objectively right or wrong. The last type of class would be more likely to have exams that would require the understanding of some concept that could be mathematically manipulated and expressed in an objective answer. These codes were determined by information available in the research itself. Each article was inspected separately on at least three different occasions to check for coding errors.

Results

A number of summary statistics are given in the tables. As pointed out by Schulze (2007), a large set of procedures has been applied to meta-analysis in the past. Cohen (1981) transformed his data by utilizing a Fisher z conversion of r, a procedure not recommended for data with large heterogeneity, such as found in this study (Hunter & Schmidt, 2004; Shulze, 2007). Table 2 summarizes the formulas utilized in the present study to provide summary statistics.

A summary of the between-class data in Table 1 shows a small positive correlation, but one in which the magnitude is not significantly different from zero. This is true irrespective of the type of summary statistic utilized. Given a random distribution of associations around zero, half would be expected to be negative and half positive. Ten of the studies found a negative association, and 32 found a positive association ($\chi^2 = 11.50$, $df = 1$, $p < .001$). Thirteen of the 42 studies (31%) found a significant positive correlation, whereas only one significant

Table 1
Summary of Learning/Student Evaluation of Teaching (SET) Studies: Between-Class Sections

Source	Ed Pub[a]	*n*	*r*	*t*[b]
Bendig (1953)	Yes	5	**.89**	3.38
Braskamp, Caulley, & Costin (1979)	Yes	19	.17	0.71
		17	**.48**	2.12
Centra (1977)	Yes	22	**.64**	3.72
		13	.23	0.78
		8	**.87**	4.32
		7	.58	1.59
		8	.41	1.10
		7	.60	1.68
		7	.61	1.72
Cohen (1981)	Yes	35	**.41**[c]	2.58
Costin (1978)	Yes	25	**.52**	2.92
		25	**.56**	3.24
		21	**.46**	2.26
		25	**.41**	2.16
Doyle & Whitely (1974)	Yes	12	.49	1.78
Frey (1973)	No	8	**.91**	5.38
		5	.60	1.30
Frey, Leonard, & Beatty (1975)	Yes	9	**.81**	3.65
		5	.74	1.91
		12	.18	0.58
Johnson (2003)	No	62	(–.11)[d]	–0.88
Palmer, Carliner, & Romer (1978)	Yes	14	(–.16)	–0.56
Rodin & Rodin (1972)	No	12	**–.75**	–3.59
Sheets, Topping, & Hoftyzer (1995)	No	58	.18[e]	1.35
		63	–.14	–1.12
Shmanske (1988)	No	17	.21	0.83
Soper (1973)	No	14	–.17	–0.60
Sullivan & Skanes (1974)	Yes	6	–.28	–0.58
		14	.42	1.60
		8	.08	0.20
		6	.55	1.32
		8	.48	1.34
		16	.34	1.35
		9	.33	0.92
		9	.57	1.84
		40	**.40**	2.69
		14	**.51**	2.05
Weinberg, Fleisher, & Hashimoto (2007)	No	194	(–.02)	–0.30
		122	(–.05)	–0.53
		88	(–.17)	–1.58
Yunker & Yunker (2003)	No	46	(–.11)	–0.73
Raw average		26.6	.33	1.28
Median		14.0	.41	1.35
Weighted average $\bar{r}$			.134	
Weighted standard error σ_e			.191	

Note: Bold type indicates that the association is significant at the .05 level.
a. Ed Pub includes publications from educational disciplines and from educational psychology.
b. *n* = number of sections; *r* = reported correlation between learning and SET; *t* = t – value of test that *r* = 0.
c. Cohen's 1981 meta-analysis contained 67 studies; those utilized in other places in this report were mathematically removed from Cohen's data. His average *r* with 67 cases was .43.
d. Regression beta coefficients were given without enough information to determine *r*. The correlation was estimated by utilizing the given *t* values.
e. The correlation is the average of 17 instructor factors; a subsequent multiple regression with numerous controls shows that this average is consistent with found probability levels.

Table 2
Date Summaries and Formulas

Statistic	Formula	Symbol	Comments
Raw averages	$\Sigma_i^k v_i / k$		k = # of studies
Weighted average	$\Sigma_i^k n_i r_i / \Sigma_i^k n_i$	$\bar{r}$	n = # of sections
Weighted error variance	$\Sigma_i^k (1 - \bar{r}^2)^2 k / \Sigma_i^k n_i$	σ_e^2	
Weighted correlation variance	$\Sigma_i^k (n_i (r_i - \bar{r}^2)^2) / \Sigma_i^k n_i$	σ_ρ^2	
Reliability (r)	$\sigma_e^2 / \sigma_\rho^2$		
Test of mediating factors beyond two variables	$\chi^2 = (\Sigma_i^k n_i / (1 - \bar{r}^2)^2)\sigma_\rho^2$		df = # of sections

Source: Hunter & Schmidt (2004); Lyons (1997).

negative correlation was noted. In general, then, there does appear to be a positive association between learning and SET, but the average association's magnitude is small and inconsistent across contingencies.

Two indicators give definitive justification for investigating the existence of mediating factors. The reliability of the learning/SET associations in this sample of studies is .403, indicating that almost 60% of the variance is unexplained. Furthermore, the χ^2 test of unexplained variance is highly significant ($\chi^2 = 104.05$, $df = 41$, $p < .001$; see Table 2 for explanations).

Table 3 shows the studies and identifies possible mediating factors. As seen in Table 4, 12 of 30 studies (40%) from educational and educational/psychology journals found significant associations, whereas 2 of 12 (17%) studies of other journals found a significant learning/SET association ($\chi^2 = 2.10$, $df = 1$, $p = .147$). No studies from business classes found a significant association, whereas 10 of 12 (83%) studies of educational/psychology classes found significant results. Math and science classes were between these two extremes, with 3 of 20 (15%) studies finding a significant association ($\chi^2 = 21.53$, $df = 2$, $p < .001$). Only 1 of 12 (8%) studies with final exams containing math applications were significant, whereas 9 of 11 studies (82%) found significant results in classes requiring no or little mathematics ($\chi^2 = 16.30$, $df = 2$, $p < .001$). Nine of 23 studies that did not utilize statistical control found a significant result, whereas no studies found a significant association if statistical control was applied to both learning and SET. The overall pattern, however, was not significant ($\chi^2 = 3.38$, $df = 2$, $p = .184$).

To investigate these relationships further, a common measure was needed to control for widely different sample sizes (from 5 to 194 sections) and differences in how the association was measured (ordinary least squares correlation, two-stage least squares, or multiple regression). It was possible in all the studies to calculate the corresponding t tests for each association (assuming $\rho = 0$). This measure utilizes an adjustment for sample size. Each t distribution, however, is distinctly platykurtic for the smaller samples found in these studies. To correct for this difference, the t values were converted to z scores that would represent the same probability of the t value but on a normal curve. For example, a correlation of 10 cases could find a $t(8) = 1.86$ with 5% of the area under the normal curve to the right. The corresponding z score would be 1.64. On the other hand, a sample of 42 studies would have a corresponding position on the t curve of $t(40) = 1.68$ but would still convert to a z value of 1.64. In other words, the Z_t score given in Table 3 is a standardized measure of the probability of finding the association given in the study, assuming the true value was equal to 0 ($\rho = 0$).

Utilizing this measure, a number of statistical tests of associational magnitude can be made (see Table 5). All of the factors identified in the study resulted in significant differences in the magnitudes of the learning/SET associations.

An inspection of the data suggests several other associations may be present. The correlation between Z_t and the size of the sample (number of sections) is negative ($r = -.373$, $p = .015$), indicating that as the size of the sample increases, the strength of the association decreases. Older studies appear to have more significant results than newer studies. The correlation between the age of the publication and Z_t was positive ($r = .482$, $p = .001$). To control for this effect, the publication age of the journal was added as a covariate in a further analysis. Sample size was controlled by the construction of the Z_t measure itself. As shown in Table 5, all the intervening factors remained significant, except for statistical controls.

The within-class data is shown in Table 6. In summary, the association is found to be very close to zero. Five of the 11 studies found a negative correlation ($\chi^2 = 0.09$, $df = 1$, $p = .764$).

Discussion

There does appear to be a small average positive association between learning and the SET in the between-class data as measured in these studies. The relationship

Table 3
Summary of Learning/Student Evaluation of Teaching (SET) Studies by Characteristics: Between-Class Sections

Source	Control[a]	Disc	Exam Objectivity[b]	Z_t
Education and educational/psychology journals				
Bendig (1953)	No	Psych	Verbal	**2.03**
Braskamp, Caulley, & Costin (1979)	No	Psych	Verbal	0.69
		Psych	Verbal	1.95
Centra (1977)	PG	Psych	Verbal	**3.20**
		Bio	Objective	0.75
		Math	Objective	**2.81**
		Physics	Object/Appl	1.36
		Chem	Object/Appl	1.01
		Chem	Object/Appl	1.43
		Bio	Objective	1.45
Cohen (1981)	—	Meta	—	**2.44**
Costin (1978)	No	Psych	Verbal	**2.66**
		Psych	Verbal	**2.92**
		Psych	Verbal	**2.11**
		Psych	Verbal	**2.04**
Doyle & Whitely (1974)	SA	French	Verbal	1.62
Frey, Leonard, & Beatty (1975)	SA	Psych	Verbal	**2.64**
		Math	Objective	**2.43**
		Math	Objective	0.56
Palmer, Carliner, & Romer (1978)	Comb	Econ	Object/Appl	–0.55
Sullivan & Skanes (1974)	No	Bio	Objective	–0.54
		Bio	Objective	1.49
		Chem	Objective	0.19
		Chem	Objective	1.13
		Math	Objective	1.22
		Math	Objective	1.28
		Math	Objective	0.86
		Physics	Object/Appl	1.60
		Psych	Verbal	**2.55**
		ElSc	Verbal	1.60
Non–educational/psychology journals				
Frey (1973)	SA	Math	Objective	**3.16**
		Math	Objective	1.07
Johnson (2003)	Comb	Mix[c]	—	–0.87
Rodin & Rodin (1972)	SA	Physics	Object/Appl	**–2.81**
Sheets et al. (1995)	No	Econ	Object/Appl	1.33
		Econ	Object/Appl	–1.11
Shmanske (1988)	PG	Econ	Object/Appl	0.81
Soper (1973)	No	Econ	Object/Appl	–0.17
Weinberg, Fleisher, & Hashimoto (2007)	Comb	Econ	Object/Appl	–0.87
		Econ	Object/Appl	–0.30
		Econ	Object/Appl	–1.55
Yunker & Yunker (2003)	Comb	Acc	Object/Appl	–0.72
Raw average				1.08
Median				1.31

Note: Bold type indicates that the association is significant at the .05 level.
a. Control = statistical control of learning and SET; No = no controls; SA = student achievement; PG = prior class grades or performance; Comb = combination of both SA and PG.
b. Type of exam: Verbal = no math; Objective = math to find answer; Object/Appl = concept problems solved with mathematics, that is, accounting or physics.
c. Johnson's data comes from a campuswide study. Of the 40 studies, 24 are from science and engineering classes, 5 are from business classes, and 16 are from language classes.

Table 4
Analysis of Associations Based on Data From Tables 1 and 3

	Type of Journal[a]		
	Education	Noneducation	Total
Not significant	18	10	28
Significant	12	2	14
Total	30	12	42

	Type of Class[b]			
	Education	Science	Business	Total
Not significant	2	17	9	28
Significant	10	3	0	13
Total	12	20	9	41

	Type of Exam[c]			
	1	2	3	Total
Not significant	2	15	11	28
Significant	9	4	1	14
Total	11	19	12	42

	Type of Statistical Control[d]			
	1	2	3	Total
Not significant	14	8	6	28
Significant	9	4	0	13
Total	23	12	6	41

a. $\chi^2 = 2.10$, $p = .147$.
b. $\chi^2 = 21.53$, $p < .001$.
c. $\chi^2 = 16.30$, $p < .001$. 1 = verbal; 2 = objective; 3 = objective/applications.
d. $\chi^2 = 3.38$, $p = .184$. 1 = no controls; 2 = partial controls; 3 = both learning and student evaluation of teaching controls.

Table 5
Analysis of Associations: Magnitude Z_t

Factor	*n*	$M(Z_t)$	*SE*
Type of journal			
Educational/psychology	30	1.56	0.17
Others	12	–0.14	0.44
$F(1, 40) = 18.87$, $p < .001$			
$F(1, 39) = 8.61$, $p = .014$ (age of publication controlled)			
Type of class (academic discipline)			
Education/psychology/language	12	2.17	0.20
Objective and science	20	0.98	0.29
Business	9	–0.31	0.30
$F(2, 38) = 13.67$, $p < .001$			
$F(1, 37) = 6.48$, $p = .004$ (age of publication controlled)			
Type of exam			
Verbal	11	2.22	0.21
Objective	19	1.24	0.24
Objective/applications	12	–0.22	0.38
$F(2, 39) = 16.98$, $p < .001$			
$F(1, 38) = 8.72$, $p = .001$ (age of publication controlled)			
Statistical controls			
None	23	1.37	0.22
Student achievement	4	0.76	1.27
Prior grades	8	1.60	0.32
Full combination	6	–0.75	0.18
$F(3, 37) = 5.92$, $p = .002$			
$F(1, 36) = 1.93$, $p = .142$ (age of publication controlled)			

is small, not universal, and subject to the following intervening variables:

1. The association is stronger in research published in educational journals than in other sources.
2. The association is strongest in studies from the education and liberal arts disciplines. There is no evidence that the association exists in business classes.
3. The type of learning measures made a difference. The more objective the measures, the smaller the learning/SET association.
4. The more statistical control was utilized to handle extraneous variables in both learning and SET, the less association was found. The attenuation of this relationship when age of publication was added as a covariant suggests simply that researchers are becoming increasingly sophisticated over time in their utilization of statistical control.
5. The findings are highly variable, even for the same researcher. Centra's (1977) correlations ranged from .23 to .87. Sullivan and Skanes' (1974) findings ranged from –.28 to .55.
6. There is no evidence in this sample that a learning/SET association exists in within-class data.

Although finding a difference between within-class and between-class data is easy to conceptualize statistically, a difference between these two measures raises difficult questions with important real-world implications. The only way that classes could show a valid learning/SET association, which was not shown by students within a class, would be if the class reached a statistical consensus and that consensus overcame individual tendencies. For example, an individual student, on average, would need to give a good instructor a higher evaluation even if that student was learning less than others and a lower evaluation to a poor instructor even if the student was learning more than others in the same class. How this consensus would be obtained and how the improbable balance between group knowledge and individual bias could be achieved is problematic.

Table 6
Summary of Learning/Student Evaluation of Teaching Studies: Within-Class Individual Student Data

Source	Ed Pub[a]	*n*	*r*/*B*[b]	*t*
Attiyeh & Lumsden (1972)	No	30,000	*–.06*	–0.57
Bendig (1953)	Yes	124	.14	1.56
Gramlich & Greenlee (1993)	No	5,066	*.02*	0.80[c]
		4,869	***.14***	2.50
		2,628	*.02*	0.60
		2,544	*.07*	0.90
		709	*–.00*	–0.01
Lundsten (1986)	No	2,069	**–.11**	–5.03
Shmanske (1988)	No	236	*.10*	1.37
Soper (1973)	No	506	*–.12*	–0.27
Yunker & Yunker (2003)	No	183	***–.20***	–2.15
Raw average				–0.03

Note: Bold type indicates that the association is significant at the .05 level.
a. Ed Pub includes publications from educational and educational psychology disciplines.
b. Values in italics are beta coefficients from regressions; roman type indicates Pearson correlations. The average association was not calculated because of these differences.
c. Gramlich and Greenlee (1993) *t* values were not given. These *t* values were estimated from significance data given in the report.

Summary and Research Implications

For almost 60 years, the debate has been about the relationship between learning and SET, which essentially is a discussion about the validity of the instruments. This literature review and meta-analysis raise the possibility that the dichotomous nature of this debate may be misplaced. The question does not appear to be *if* a relationship exists, but rather *when* the relationship exists.

Any global explanation advanced to explain the contradictions and patterns identified in the literature review and meta-analysis would need to reconcile five conclusions. First, the students' perceptions of their grades appear to be related to the evaluations they give both for the class and the instructor. Both a leniency and a reciprocity effect have been found. The research seems to suggest that these effects will be modified by a variety of influences irrespective of "learning." It has been suggested that because business and marketing classes typically have lower average grades than many humanity and education courses, they would therefore be expected to show a larger grade/evaluation association simply as a statistical artifact (Clayson, 2007). It is important to keep in mind that the grade/evaluation association is not equivalent to a learning/SET relation.

Second, the research almost universally finds a negative association between rigor and learning on the SET. Students seem to associate rigor with negative instructor characteristics that override positive learning relationships.

Third, the more objective the measurement process becomes, both for learning and SET, the more the learning/SET association is reduced. More recent and larger studies (Johnson, 2003; Weinberg et al., 2007) have found that although a positive relationship exists between student perceptions of learning and the evaluation, that relationship cannot be found when more objective measures of learning are utilized. As statistical sophistication has increased over time, the reported learning/SET relationship has generally become more negative. Almost 40 years have passed since the positive result in Sullivan and Skanes' (1974) study was obtained. No study could be found after 1990 that showed a positive significant relationship between learning and the SET. How learning is operationalized appears to have important implications. In some disciplines, grades are typically derived from work that has an objective "correct" answer, as opposed to measures in a more subjective field of study. Some believe that classes utilizing quantitative applications typically require more cognitive skills during testing than do areas that may rely more on measures emphasizing memorized feedback (McKeachie, 1987). The Rodin and Rodin (1972) study, which found the largest negative association between learning and SET, not only tested understanding, but also tested it at the paradigm level. The largest positive correlation (.91) was found by Frey et al. (1975), who utilized a relatively small sample of introductory psychology courses.

Fourth, due to discipline differences the meta-analysis indicates that the academic discipline area is an important variable. This has been overlooked in much of the literature. Although there has been a careful attempt to control for extraneous variables, the academic areas of the classes are not always clarified in SET studies.

Fifth, some differences in findings appear real, and not entirely artifacts of differing methodology. There is little evidence to suggest that all differences are solely situational or due to methodology. For example, there are no mistakes large enough in Cohen's meta-analysis to invalidate his conclusion that learning is positively related to SET. On the other hand, there are no errors found in Johnson's (2003) or Weinberg and colleagues' (2007) work that could logically be said to invalidate their conclusions that learning and SET are not related or may even be negatively related.

Summary Explanation

These five conclusions lead the writer to advance the following summary explanation:

> Objective measures of learning are unrelated to the SET. However, the students' satisfaction with, or perception of, learning is related to the evaluations they give.

This explanation suggests that the validity of the relationship between learning and SET is situational and that the associations, and the validities of those associations, depend on a number of factors. To a certain extent, the explanation can be summed up by a rather dark statement about human behavior by the American journalist and author Donald R. P. Marquis, who once wrote, "If you make people think they're thinking, they'll love you. If you really make them think, they'll hate you" (as cited in Morley & Evertt, 1965, p. 237). In summary, the learning/SET association is valid to the extent that the student's perception of learning is valid. The literature, however, indicates that students do not always hold a realistic evaluation of their own learning. This allows for a number of predictions that could be clarified by future research.

Suggested Research

The meta-analysis discussed in this article suggests four hypotheses for further investigation. First, instructors' ability to convince their students that they are learning will be related to the evaluations. This prediction is consistent with the very large association found between instructor personality and SET (see Clayson & Sheffet, 2006, for a review of personality issues in SET).

Second, the more objective a measure of learning becomes, the more negative that measure is when associated with SET. Studies utilizing student perception have generally found positive associations between learning and the evaluations. When learning has been defined in more objective terms, removed from the students' and/or instructors' own subjective interpretations, the correlation tends to fall into nonsignificant or even into negative ranges.

Third, the association between learning and SET depends on what cognitive skills are utilized to create a measure of learning. The meta-analysis suggests that the association may be related to the type of learning being measured. On one extreme, learning could be measured by highly subjective evaluations, typically utilizing only subjective feelings or memorization skills with which the students have years of experience. The opposite extreme would be characterized by concept learning, analytical skills, and abstraction, leading to an objective measure of learning. The learning/SET association would be expected to run from positive to neutral down this continuum.

Fourth, the validity of the SET instruments for any given instructor will depend on whether instructional and institutional goals related to learning are consistent with student perceptions. Holding instructor personality constant, the evaluations could be a valid measure of how good an instructor teaches if the class is graded in a fashion consistent with institutional instructional goals that are also consistent, for whatever reason, with students' perceptions of their own learning. If, however, students are graded using criteria, either objective or subjective, that violate their own perceptions of learning, then the validity of the instruments will become questionable. If true, this hypothesis would suggest the evaluations do not have equal validity across faculty, class topic matter, and academic disciplines.

Consequences for Marketing Education

Assuming the conclusions of this review are correct, we would find inconsistent associations between learning and SET in marketing classes. We might expect, for example, to find no or even a slight negative correlation between evaluations and learning in classes such as marketing research, especially if the course is highly structured and statistical in nature. As Johnson (2003) found in such classes, instructors with low evaluations may actually be producing students who perform better in subsequent courses. A class such as advertising, if the students were more subjectively graded, may show a positive relationship between learning and SET. Consumer behavior might show any combination of associations, depending on the instructor's approach and how the students' grades were derived. If test material attempted to measure understanding and application of theoretical concepts, we could expect to find no associations between evaluations and learning. On the other hand, if testing was based mostly on recognition and memorization, we would expect the association to become more positive.

SET would be more valid for a marketing instructor teaching a class that is graded in a fashion consistent with students' perceptions of their own learning. On the other hand, business classes that contain math applications and require more cognitive skills for students to demonstrate their level of competency may find the SET to be relatively invalid. Instruction in these classes may actually be harmed if too much attention is paid to the instructor's evaluations.

This would imply that on balance, universal and group-weighted SET results should not be utilized, or they should be interpreted with great care. First, little information may be obtained about how much students are learning by looking only at the evaluations. Second, instructors who are teaching students to think, and to stretch mentally and professionally, could actually be penalized. As summarized by Paswan and Young (2002) at the conclusion of their study of business students, "Instead of asking instructors to improve teaching evaluations, schools should be asking themselves whether they should be asking instructors to make the course more or less demanding, interactive, or structured and organized" (p. 200).

References

Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research*, *58*(2), 151-179.

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

Adams, J. B. (2005). What makes the grade? Faculty and student perceptions. *Teaching of Psychology*, *32*(1), 21-24.

Adams, J. V. (1997). Student evaluations: The rating game. *Inquiry*, *1*(2), 10-16.

Aleamoni, L. M. (1999). Student ratings myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153-166.

Attiyeh, R., & Lumsden, K. G. (1972). Some modern myths in teaching economics: The U. K. experience. *American Economic Review*, *62*, 429-433.

Bacon, D. R., & Novotny, J. (2002). Exploring achievement striving as a moderator of the grading leniency effect. *Journal of Marketing Education*, *24*, 4-14.

Baird, J. S. (1987). Perceived learning in relation to student evaluation of university instruction. *Journal of Educational Psychology*, *79*, 9091.

Bendig, A. W. (1953). The relationship of level of course achievement to students' instructor and course ratings in introductory psychology. *Educational and Psychological Measurement*, *13*, 437-448.

Bharadwaj, S., Futrell, C. M., & Kantak, D. M. (1993). Using student evaluations to improve learning. *Marketing Education Review*, *3*(2), 16-21.

Birnbaum, M. H. (2000). *A survey of faculty opinions concerning student evaluation of teaching*. Retrieved June 21, 2008, from http://psych.fullerton.edu/mbirnbaum/faculty3.htm

Boex, L. F. J. (2000). Attributes of effective economics instructors: An analysis of student evaluations. *Research in Economic Education*, *31*, 211-227.

Braskamp, L. A., Caulley, D., & Costin, F. (1979). Student ratings and instructor self-rating and their relationship to student achievement. *American Educational Research Journal*, *16*, 295-306.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performances*. San Francisco: Jossey-Bass.

Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research* (IDEA Paper No. 20). Manhattan: Center for Faculty Evaluation & Development, Division of Continuing Education, Kansas State University.

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

Centra, J. A. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal*, *14*, 17-24.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, *44*, 495-518.

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

Chonko, L. B., Tanner, J. F., & Davis, R. (2002). What are they thinking? Students' expectations and self-assessments. *Journal of Education for Business*, *77*, 271- 281.

Clayson, D. E. (1994). Contrasting results of three methodological approaches on the interpretation of a student evaluation of instruction. In E. W. Chandler (Ed.), *Proceedings of the Midwest Marketing Association* (pp. 209-214). Chicago: Midwest Marketing Association.

Clayson, D. E. (2004). A test of the reciprocity effect in the student evaluation of instructors in marketing classes. *Marketing Education Review*, *14*(2), 11-21.

Clayson, D. E. (2005a). Performance overconfidence: Metacognitive effects or misplaced student experience. *Journal of Marketing Education*, *27*, 122-129.

Clayson, D. E. (2005b). Within-class variability in student-teacher evaluations: Example and problems. *Decision Sciences Journal of Innovative Education*, *3*(1), 109-124.

Clayson, D. E. (2007). Conceptual and statistical problems of using between-class data in educational research. *Journal of Marketing Education*, *29*, 34-38.

Clayson, D. E., Frost, T. F., & Sheffet, M. J. (2005). Grades and the student evaluation of instruction: A test of the reciprocity effect. *Academy of Management Learning & Education*, *5*(1), 52-65.

Clayson, D. E., & Haley, D. A. (1990). Student evaluations in marketing: What is actually being measured? *Journal of Marketing Education*, *12*, 9-17.

Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education*, *28*, 149-160.

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

Comm, C. L., & Manthaisel, D. F. X. (1998). Evaluating teaching effectiveness in America's business schools: Implications for service marketers. *Journal of Professional Service Marketing*, *16*(2), 163-170.

Costin, F. (1978). Do student ratings of college teachers predict student achievement? *Teaching of Psychology*, *5*(2), 86-88.

Cruse, D. B. (1987). Student evaluations of the university professor: Caveat professor. *Higher Education*, *16*, 723-737.

Dowell, D. A., & Neal, J. A. (1982). A selective review of the validity of student ratings of teaching. *Journal of Higher Education*, *53*, 51-62.

Doyle, K. O., & Whitely, S. E. (1974). Student ratings as criteria for effective teaching. *American Educational Research Journal*, *11*, 259-274.

Faranda, W. T., & Clarke, I., III. (2004). Student observations of outstanding teaching: Implications for marketing educators. *Journal of Marketing Education*, *26*, 271-281.

Flowers, L., Osterlind, S. J., Pascarella, E. T., & Pierson, C. T. (2001). How much do students learn in college? *Journal of Higher Education*, *72*, 565-583.

Frey, P. W. (1973). Student ratings of teaching: Validity of several rating factors. *Science*, *182*(4107), 83-85.

Frey, P. W., Leonard, D. W., & Beatty, W. W. (1975). Student ratings of instruction: Validation research. *American Educational Research Journal*, *12*, 435-447.

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

Gaultney, J. F., & Cann, A. (2001). Grade expectations. *Teaching of Psychology*, *28*(2), 84-87.

Gillmore, G. M., & Greenwald, A. G. (1999). Using statistical adjustment to reduce biases in student ratings. *American Psychologist*, *54*, 518-519.

Goldberg, G., & Callahan, J. (1991). Objectivity of student evaluations of instructors. *Journal of Education for Business*, *66*, 377-378.

Goldman, L. (1985). The betrayal of the gatekeepers: Grade inflation. *Journal of General Education*, *37*(2), 97-121.

Gramlich, E. M., & Greenlee, G. A. (1993). Measuring teaching performance. *Research in Economic Education*, *24*(1), 3-13.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, *52*, 1182-1186.

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

Greenwald, A. G., & Gillmore, G. M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology*, *89*, 743-751.

Gremler, S. D., & McCollough, M. A. (2002). Student satisfaction guarantees: An empirical examination of attitudes, antecedents, and consequences. *Journal of Marketing Education*, *24*, 150-160.

Grimes, P. W. (2002). The overconfident principles of economics students: An examination of metacognitive skill. *Journal of Economic Education*, *33*(1), 15-30.

Guthrie, E. R. (1954). *The evaluation of teaching: A progress report*. Seattle: University of Washington.

Hake, R. R. (2002). *Problems with student evaluations: Is assessment the remedy?* Retrieved June 19, 2007, and February 13, 2008, from http://physics.indiana.edu/~hake/assesstherem1.pdf

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

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousands Oaks, CA: Sage.

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York: Springer.

Kaplan, M., Mets, L. A., & Cook, C. E. (2000). *Questions frequently asked about student ratings forms: Summary of research findings*. Retrieved March 31, 2006, from http://www.crlt.umich.edu/tstrategies/studentratingfaq.html

Kennedy, E. J., Lawton, L., & Plumlee, E. L. (2002). Bliss ignorance: The problem of unrecognized incompetence and academic performance. *Journal of Marketing Education*, *24*, 243-252.

Kolevzon, M. S. (1981). Grade inflation in higher education: A comparative study. *Research in Higher Education*, *15*, 195-212.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research*, *109*, 9-25.

Laverie, D. A. (2002). Improving teaching through improving evaluation: A guide to course portfolios. *Journal of Marketing Education*, *24*, 104-113.

Lundsten, N. L. (1986). Student evaluations in a business administration curriculum: A marketing viewpoint. *AMA Developments in Marketing Science*, *9*, 169-173.

Lyons, L. C. (1997). *Meta-analysis: Methods of accumulating results across domains*. Retrieved February 6, 2008, from http://www.lyons-morris.com/MetaA/index.htm

Machina, K. (1987). Evaluating student evaluations. *Academe*, *73*(3), 19-22.

Marks, R. B. (2000). Determinants of student evaluations of global measures of instructor and course value. *Journal of Marketing Education*, *22*, 108-119.

Marlin, J. W., & Niss, J. F. (1980). End-of-course evaluations as indicators of student learning and instructor effectiveness. *Journal of Economic Education*, *11*(2), 16-27.

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

Marsh, H. W., Hau, K., Chung, C., & Siu, T. L. (1997). Students' evaluations of university teaching: Chinese version of the Students' Evaluations of Educational Quality instrument. *Journal of Educational Psychology*, *89*, 568-572.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist*, *52*, 1187-1197.

Marsh, H. W., & Roche, L. A. (1999). Reply upon SET research. *American Psychologist*, *54*, 517-518.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, *92*, 202-228.

McKeachie, W. J. (1987). Commentary: Instructional evaluation: Current issues and possible improvements. *Journal of Higher Education*, *58*, 344-350.

Moore, M., & Trahan, R. (1998). Tenure status and grading practices. *Sociological Perspectives*, *41*, 775-781.

Moreland, R., Miller, J., & Laucka, F. (1981). Academic achievement and self -evaluations of academic performances. *Journal of Educational Psychology*, *73*, 335-344.

Morley, C., & Evertt, L. D. (Eds). (1965). *Bartlett's familiar quotations*. New York: Pocket Books.

Morsh, J. E., Burgess, G. G., & Smith, P. N. (1956). Student achievement as a measure of instructor effectiveness. *Journal of Educational Psychology*, *47*(2), 79-88.

Palmer, J., Carliner, G., & Romer, T. (1978). Leniency, learning, and evaluations. *Journal of Educational Psychology*, *70*, 855-863.

Paswan, A. K., & Young, J. A. (2002). Student evaluation of instructors: A nomological investigation using structural equation modeling. *Journal of Marketing Education*, *24*, 193-202.

Pollio, H. R., & Beck, H. P. (2000). When the tail wags the dog. *Journal of Higher Education*, *71*, 84-102.

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

Redding, R. E. (1998). Students' evaluation of teaching fuel grade inflation. *American Psychologist*, *53*, 1227-1228.

Remmers, H. H., & Brandenburg, G. C. (1927). Experimental data on the Purdue Rating Scale for Instruction. *Educational Administration and Supervision*, *13*, 519-527.

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

Rosenthal, R., Rosnow, R. I., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research*. Cambridge, UK: Cambridge University Press.

Ryan, J. J., Anderson, J. A., & Birchler, A. B. (1980). Student evaluation: The faculty responds. *Research in Higher Education*, *12*, 317-333.

Schlee, R. P. (2005). Social styles of students and professors: Do students' social styles influence their preferences for professors? *Journal of Marketing Education*, *27*, 130-142.

Schmidt, T. A., Houston, M. B., Bettencourt, L. A., & Boughton, P. D. (2003). The impact of voice and justification on students' perceptions of professors' fairness. *Journal of Marketing Education*, *25*, 177-186.

Schulze, R. (2007). Current methods for meta-analysis: Approaches, issues, and developments. *Zeitschrift für Psychologie/Journal of Psychology*, *215*(2), 90-103.

Schwab, D. P. (1976). *Manual for the Course Evaluation Instrument*. Madison: University of Wisconsin, School of Business.

Scriven, M. (1983). Summative teacher evaluations. In J. Milman (Ed.), *Handbook of teacher evaluation* (pp. 244-271). Thousand Oaks, CA: Sage.

Seiver, D. A. (1983). Evaluations and grades: A simultaneous framework. *Journal of Economic Education*, *14*(3), 32-38.

Seldin, P. (1993, July 21). The use and abuse of student ratings of professors. *Chronicles of Higher Education*, *39*(46), A40.

Seldin, P. (1999). *Changing practices in evaluating teaching: A practical guide to improving faculty performance and promotion/tenure decisions*. Bolton, MA: Anker.

Sheehan, E. P., & DuPrey, T. (1999). Student evaluations of university teaching. *Journal of Instructional Psychology*, *26*, 188-193.

Sheets, D. F., Topping, E. E., & Hoftyzer, J. (1995). The relationship of student evaluations of faculty to student performance on a common final examination in the principles of economics courses. *Journal of Economics*, *21*(2), 55-64.

Shmanske, S. (1988). On the measurement of teacher effectiveness. *Research in Economic Education*, *19*, 307-314.

Simpson, P. M., & Siguaw, J. A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education*, *22*, 199-213.

Sixbury, G. R., & Cashin, W. E. (1995). *Description of database for the IDEA Diagnostic Form* (IDEA Technical Report No. 9). Manhattan: Center for Faculty Evaluation & Development, Division of Continuing Education, Kansas State University.

Soper, J. C. (1973). Soft research on a hard subject: Student evaluations reconsidered. *Journal of Economic Research*, *5*(1), 22-26.

Stapleton, R. J., & Murkison, G. (2001). Optimizing the fairness of student evaluations: A study of correlations between instructor excellence, study production, learning production, and expected grades. *Journal of Management Education*, *25*, 269-291.

Steiner, S., Holley, L. C., Gerdes, K., & Campbell, H. E. (2006). Evaluating teaching: Listening to students while acknowledging bias. *Journal of Social Work Education*, *42*, 355-376.

Stumpf, S. A., & Freedman, R. D. (1979). Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology*, *71*, 293-302.

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

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research*, *27*(5), 45-56.

Weinberg, B. A., Fleisher, B. M., & Hashimoto, M. (2007). *Evaluating methods for evaluating instruction: The case of higher education* (NBER Working Paper No. 12844). Retrieved June 21, 2008, from http://www.nber.org/papers/w12844

Wilhelm, W. B. (2004). The relative influence of published teaching evaluations and other instructor attributes on course choice. *Journal of Marketing Education*, *26*, 17-30.

Williams, W. M., & Ceci, S. J. (1997). "How'm I Doing?": Problems with student ratings of instructors and courses. *Change*, *29*(5), 13-23.

Wilson, R. (1998). New research casts doubt on value of student evaluations of professors. *Chronicle of Higher Education*, *44*(19), A12-A14.

Yunker, P. J., & Yunker, J. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business*, *78*, 313-317.

Dennis E. Clayson is a professor of marketing at the University of Northern Iowa.