JET
The Journal of Effective Teaching
an online journal devoted to teaching excellence

What Makes a Good Criminal Justice Professor? A Quantitative Analysis of Student Evaluation Forms

Patrick M. Gerkin[1] and Christopher A. Kierkus
Grand Valley State University, Grand Rapids, MI, 49508

Abstract

The goal of this research is to understand how students define teaching effectiveness. By using multivariate regression analysis of 8,000+ student evaluations of teaching compiled by a School of Criminal Justice at a Midwestern public university, this paper explores the relationships between individual indicators of instructor performance (e.g. clarity of the instructor, concern for student progress etc.), and overall impressions of instructor quality. Further consideration is given to factors like basic student characteristics (e.g. academic status), basic course characteristics (e.g. whether the course is a research methods class), basic instructor characteristics (e.g. instructor gender), and the expected grade in the class. Results indicate that the individual instructor performance items are the most important indicators of overall perceived quality. Other factors do matter, but their influence is statistically weaker. Interestingly, once the individual instructor performance items are controlled, the effect of perceived grade is rendered statistically insignificant.

Keywords: Teaching effectiveness, student evaluations.

Student evaluations of teaching (SETs) appear to be a topic of never-ending debate and inquiry. Research has been accumulating for decades about the value and validity of student's evaluation of faculty performance in the classroom. This line of inquiry began in 1927 with the first published research on the topic of student ratings (Remmer & Brandenbur, 1927). Student evaluations of teaching at the time were in their infancy as the first record of student evaluations occurred at the University of Washington only a year or two before (Guthrie, 1954). Despite intensive debate over their value, the use of SETs has grown almost continuously at universities across the United States (Marsh, & Roche, 1997; Seldin, 1993).

The ferocity of the debate, and continued scholarly inquiry, are not too surprising given that as the usage of SETs has expanded, so too has their utilization. Student evaluations of teaching are now considered in numerous settings, and at various levels within the university. Today, it is common for SETs to be requested from incoming applicants to be used in the hiring process; they are used at many institutions throughout the tenure and promotion process; and they may also be used as a means of evaluating candidates for teaching awards. Most certainly, faculty use them for formative purposes, to evaluate

[1] Corresponding author's email: gerkinp@gvsu.edu

and assess their own performance and to guide their personal efforts to improve their teaching.

While the use of SETs for formative purposes is not particularly controversial, their use as a summative measure, designed to assist in personnel decisions has led to significant objection. Formative evaluations have, in fact, been found to provide useful feedback and may improve undergraduate education (Griffin & Pool, 1998; Marsh & Roche, 1997). The controversy is generally centered around two major issues (1) the validity and reliability of instructional evaluations by students and (2) the objectives and possible applications of such evaluations (Miron & Segal, 1986). Additionally, it should be noted that research conducted by Franklin and Theal (2001) demonstrates a lack of knowledge about the literature pertaining to student ratings, and lack of statistical expertise among the faculty and administrators who are assigned to interpret them. “Even when the data are technically rigorous, one of the major problems is day-to-day practice: student ratings are often misinterpreted, misused, and not accompanied by other information that allows users to make sound decisions” (Franklin & Theal, 2001, p. 46).

The debate over the value, validity and reliability of SETs is further intensified by programs like those implemented at Texas A&M, where according to Katherine Mangan (2009), the school has started awarding bonuses ranging from $2,500 to $10,000 to faculty members who receive the highest SETs at the end of the term. Although critics of this program fear that it will lead to grade inflation as professors strive to win the admiration of their students, Mangan (2009) notes that a similar program was already in place at the neighboring University of Oklahoma. She goes on to describe how another Texas institution, the Acton School of Business, ties instructors pay even more closely with student evaluations. The faculty at this small private institution are paid $5,000 per course, with the remainder of their salary determined solely by evaluations.

As the debate over the value and proper use of student evaluations continues, SETs continue to be utilized for a variety of purposes. This study seeks to make a significant contribution to the scholarship of teaching and learning by offering findings from a large sample of student evaluations compiled by a school of criminal justice at a major university, to determine what variables are most important to students when evaluating the overall quality of their teachers. While there has been much research focused on the validity and potential biases in student evaluations of teaching, far less attention has been given to the significance of individual items in a SET, and their relationship to students’ overall perceptions of teacher effectiveness.

Literature Review

No matter their utilization, or actual value, SETs are clearly having an impact on faculty performance in the classroom. One survey of faculty found 70% of professors believe that their grading leniency, and course difficulty, bias student ratings, and 83% admitted making their courses easier in response to student evaluations (Ryan, Anderson & Birchler, 1980). This suggests whether one believes that SETs are valuable, for formative or summative purposes, their use clearly influences faculty behavior in the classroom.

The research included here will provide an overview of significant issues related to SETs, including research devoted to the issues of validity and reliability, instructor attributes such as race and gender, the influence of grades, course characteristics such as class size and type of course, and instructor qualities like clarity, concern for student progress, and fairness.

Validity

Hundreds of articles have been written on the validity of SET ratings, and it seems that most researchers agree that validating SETs is no easy task. Marsh and Roche (1997) note that SETs are difficult to validate because there is no single criterion of effective teaching. Kulik (2001) agrees:

> If ratings are valid, students will give good ratings to effective teachers and poor ratings to ineffective ones. The size of the correlation between ratings and effectiveness will provide a precise index of the validity of student ratings. The catch is that no one knows what measure to use as the criterion of teaching effectiveness. Researchers have long searched for the perfect criterion (p. 10).

Despite a failure to find the *perfect* criterion, several criteria have emerged, and appear frequently in the literature. These include measures of student learning, alumni assessments of teachers, and classroom observations by experts. Even these criteria appear less than perfect. Scriven (1983) offers a clear articulation of the inadequacies of these measures.

Given the obstacles identified here, numerous researchers have turned to an alternative construct-validation approach that seeks to relate SET ratings with a wide variety of indicators of effective teaching (Howard, Conway, & Maxwell, 1985; Kulik, 2001; Marsh, 1987). A list of the common indicators includes those previously identified as less than perfect measures when used individually. The idea is that together, they can offer a validation of SETs.

Drawing from the expansive literature on student evaluations of teaching, it is possible to offer a few general conclusions about which there appears to be some consensus. Kulik's (2001) meta-analysis of the validity and utility of student ratings concludes that "student ratings agree well with other measures of teaching effectiveness; learning measures, student comments, expert observations, and alumni ratings" (p. 23). Moreover, multisection validity studies have been able to demonstrate that SET ratings do reflect the level of student learning; which is an important criterion of effective teaching. Multisection validity studies collect data from multiple sections of the same course, taught by different professors, who administer the same standardized final exam. Two meta-analyses of such studies (Marsh, 1987; Marsh & Dunkin, 1992) suggest that the sections with the highest SETs are also the sections that perform best on the standardized final examination. Cohen (1981), in a widely cited article on the relationship between SET ratings and learning, found significant correlations between ratings and student performance on a standardized final exam. Finally, Steiner, Holley, Gerdes, and Campbell (2006) found that how much students perceive they learned in a course is an important predictor of SET scores.

However, not all validation measures have proven so successful. Measures of teacher effectiveness by colleagues and administrators, after classroom visitations, are generally unreliable and not systematically correlated with SETs, or other indicators of effective teaching (Centra, 1979; Marsh 1987; Murray, 1980).

Furthermore, it is safe to say that the issue of validity is far from settled. In the following sections the authors discuss some of the major variables that potentially bias student perceptions of faculty teaching.

Reliability

On the question of whether SETs provide reliable measures of teaching effectiveness, there is now a substantial body of evidence: according to Theal and Franklin (2001), "Whether reliability is measured within classes, across classes, over time, or in other ways, student ratings are remarkably consistent" (p. 50).

Marsh (1987) offers extensive evidence of SET reliability both over time, and across various courses, when taught by the same professor. He concludes that "given a sufficient number of students, the reliability of student evaluations compares favorably with that of the best objective tests" (p. 9). Marsh also summarizes the findings of a longitudinal study where a panel of students was asked to evaluate a course at the end of the term, and again several years later. The end of term evaluations of 100 courses correlated .83 with the ratings provided several years later. The median ratings were nearly identical.

Finally, Marsh reports his findings from data collected on more than 1,000 courses including sections of the same course, taught by the same instructor, different courses taught by the same instructor, and the same course taught by different instructors. His analysis showed that student ratings are a significantly better reflection of the effectiveness of the teacher over the influence of the course. When the same instructor taught the same course the correlation between overall instructor ratings was (.72). When the same instructor taught different courses, the correlation was (.61). Finally, when the same course was taught by a different instructor, the correlation was (-.05).

Race

There is a small, but growing, number of studies that have examined the effect of instructor race on SET ratings. This issue, while certainly significant in and of itself, may become more important given that the number of African Amercians pursuing graduate studies in the United States has risen dramatically (Gabbidon, 2002). Inevitably, this will lead to an increase in the number of African American faculty.

Several recent studies have found that race does matter in SETs. Hendrix (1998) collected data through observation, semi-structured interviews, and open-ended interviews with six professors (three black and three white) and 28 students. The purpose of the research was to explore how race affects an instructors' credibility. When interviewed, students suggested that they apply "more stringent credibility standards to professors depending on a combination of professor race and subject matter" (p.748). Interestingly, Hendrix also found that once African American professors establish credibility, students generally hold favorable attitudes.

Hase (2001) discovered that being a foreign born instructor carried with it both advantages and disadvantages. In recounting her years of teaching American students, Hase describes how her ancestry provided her with instant credibility if the course was in some way related to her identity. However, she also had negative experiences where her identity led to what she perceived as undue criticism. As one example, Hase offers a specific example from one course on global gender issues. In this course, Hase's critical examination of the United States, within the context of the subject matter, resulted in negative experiences that she attributed to being a foreign born instructor.

A more comprehensive study of the effects of race on SETs was conducted by Gabbidon (2002). He compared student evaluations from research methods courses taught at two different institutions. Noting that there was little to no variation in the style and content of the course over the period under study, the major difference between the courses was that some classes were taught at a historically black institution (98% minority students) and the other courses were taught at a primarily white institution. Gabbidon (2002) compared two specific items from the institutions SETs that assessed the overall quality of the instructor. The findings indicated no significant differences between the ratings of the students at the two universities. Gabbidon (2002) also compared the items from the SETs from a race related criminal justice course. Although, significant differences did emerge, he cautions that student interest in the topic, or the subject matter itself, could have produced the differences observed.

The findings of each of these studies on the significance of race in SETs is limited. As Gabbidon (2002) and others acknowledge, their findings are based upon the experiences of only single instructors. Nonetheless, they do address important questions about the influence of race on students' evaluation of their faculty.

Gender

The research focused on the effects of gender in SETs has been mixed. On the one hand, numerous studies have found that female instructors receive lower evaluations than their male peers (Neath 1996; Sandler 1991; Lueck 1993). In fact, Neath (1996), in an article appropriately titled *How to Improve Your Teaching Evaluations Without Improving Your Teachings* suggests that changing one's gender (if female) will improve one's effectiveness ratings. However, research conducted by Rowden & Carlson (1996) found that female instructors were rated significantly higher than their male counterparts. These studies demonstrate the lack of consensus about this issue, a point echoed by Marsh and Roche (1997), whose review of research on this topic found inconsistent findings, ultimately leading the authors to conclude that instructor gender has little to no effect on overall student ratings.

Grades

One is not likely to find a more controversial issue with regard to the value of SETs than the issue of grades. The oft-heard hypothesis is that inflating grades is the surest way to positive SETs. There is in fact some evidence to suggest that this is true. A number of studies indicate that giving out higher grades translates into higher SETs (Chako, 1983; Neath, 1996; Lersch and Greek, 2001). In fact, the research conducted by Lersch and Greek (2001) suggests a strong association between grades and SET scores. Their study examined course evaluation data from criminology and criminal justice programs at three first tier research universities in the state of Florida. The evaluations of these institutions were the same due to a requirement by the Florida Board of Regents that requires all state universities to use the same evaluation form. In their analysis, grades were the strongest correlate of instructor effectiveness. Although, the researchers go on to note that other significant variables such as the students major, level of interest, perceived fairness of the grading system, and motivation were not included in the study.

Despite these findings, the strength of the relationship between grades and SETs remains in question. A meta-analysis by Ellis (1985) found only low correlations between grades and overall perceived teaching quality. Similarly, Eison (1999), who found a statistically significant correlation between grades and student evaluation reports in 28 out of 50 studies, noted that the mean correlation was relatively weak (.14). Finally, Steiner et al. (2006) note that research consistently suggests that the grade a student believes he/she will receive in a class influences student responses to SETs. Generally speaking, it appears that the overall effect is small, but the nature of the relationship remains poorly understood.

Class Size and Type of Course

Researchers have also recognized that variables aside from individual teacher characteristics and grades can influence SETs. In fact, some of these variables have to do with the type and size of the course, variables over which faculty usually have little control. Two such examples that will be discussed here: class size and the type of course being offered.

Research devoted to the effects of class size on SETs has been mixed. While there is some evidence to suggest that smaller classes produce higher SETs, these results are far from conclusive. Gleason (1986) examined 30 studies and found that 20 concluded that smaller class sizes received higher SET scores. He went on to report that even where class size demonstrated a statistically significant effect, the effect was modest.

The type of course (e.g. required or elective) has been a variable of interest in numerous studies. A meta-analysis conducted by Neath (1996) indicated that required courses are bad for SET scores. According to Neath, required courses, "although educationally important, tend not to be well-liked, even by students within the major" (p. 1367). Marsh and Roche (1997) agree. Their findings suggest that higher evaluations are more common when the reason for taking the course is general interest, or major elective, as opposed to major requirement, or general education requirement.

Individual Measures of Instructor Performance

Finally, there are a number of studies that examine individual teacher performance items in relation to overall SET scores. In 1976 Feldman found that five dimensions of teaching had the highest correlations with actual overall evaluation. These dimensions include: stimulation of interest, clarity of explanation, intellectual challenge, sensitivity to class level and progress, and preparation or organization.

More recently Spencer and Schmelkin (2002) found that a group of sophomores, juniors, and seniors attending a private university defined effective teaching as concern for students, valuing student opinions, clarity in communication, and openness toward varied opinions. In 2005, Okpala and Ellis collected data from 218 different colleges in the U.S. regarding student perceptions of teacher quality. Their analysis uncovered five important qualities of effective teaching including caring for students and their learning, content knowledge, verbal skills, dedication to teaching, and teaching skills.

Some researchers have included college faculty themselves in their inquiries into what makes a quality teacher. Schaefer, Epting, Zinn and Buskit (2003) asked 99 faculty, and 231 students, to identify the top 10 qualities of effective college teachers from a list of 28 qualities. Although the orders did vary, both groups agreed upon eight of the top 10 traits: approachable, creative and interesting, encouraging and caring, enthusiastic, flexible and open minded, knowledgeable, realistic expectations and fair, and respectful.

Statement of Research Problem

The general goal of this project is to identify the characteristics that make a good college professor. More specifically, our intent is to explore a series of questions drawn from student evaluation forms, and to determine which of the individual items are most strongly tied to overall student perceptions of instructor effectiveness. The items evaluated in this analysis range from general demographic indicators (e.g. student and instructor gender), to class characteristics (e.g. whether the student had a strong desire to take the course, whether the class was in the student’s intended major), to expected course grades, and more specific measures of instructor performance (e.g. whether the instructor was well prepared for class, whether he or she demonstrated concern for student progress).

Sample and Data

This research represents a secondary analysis of data collected as part of the teacher evaluation process at a Midwestern university. This university is a Master’s degree granting institution with a total enrollment ranging between approximately 23,000 and 24,000 students over the study period. The School of Criminal Justice (SCJ) grants both Bachelors and Masters degrees in criminal justice, and also features a legal studies program and a police academy. The annual enrollment of the SCJ was between 700 and 800 students during the period studied. Individual undergraduate classes ranged between a high of 100-150 students for a large section of CJ 101 (justice in society / introduction to criminal justice) to fewer than 20 in research methods classes such as CJ 300, or the senior capstone course (CJ 495).

As part of one of the class meetings held during the final few weeks of the semester, each instructor is expected to distribute course evaluation forms to his or her students. The forms have both a quantitative component (where students respond to series of close ended questions) and a qualitative component (where they can provide open ended comments). While the exact procedure for administration of the evaluations varies from instructor to instructor, students are typically given approximately 15 minutes at the end of a class to complete the survey. The instructor leaves the room during the evaluation period, and the forms are returned to the instructor's department in a sealed envelope.

For the purpose of this analysis, the authors compiled a sample of all of the student evaluation surveys completed by pupils enrolled in 5 specific courses (CJ 101, 201, 300, 312 and 495 - see Table 1 for course descriptions), taught by instructors in the School of Criminal Justice between the years 2004 and 2009. Over the study period, the SCJ employed approximately 15 full-time, tenured, or tenure track faculty, and a varied number of visiting and adjunct professors. Because the data was identity stripped by both student and instructor, it was not possible for the authors to determine the identity of any faculty. However, in general, over 2/3 of the courses in the SCJ are taught by full time professors: the only class that was likely to have been taught by a part time faculty member on a regular basis would have been selected sections of CJ 101.

A total of 8,117 completed surveys were used to construct the data set for this project. Table 1 presents a description of each course, and the number of surveys drawn from those classes. These five specific courses were chosen because they represent a broad spectrum of criminal justice topics, and because the SCJ teaches multiple sections of each course, to a substantial number of students, every year. This minimized the chances that any individual instructor, or student, could be inadvertently identified as part of this project. Hence, although the sample in question can most accurately be characterized as convenience based, it is likely to be broadly representative of the experiences of criminal justice students, and instructors, at mid-sized Midwestern universities. The size of the sample (n > 8,000) also permits substantial statistical power for quantitative analysis.

Variables

Independent Variables

In total, the analysis utilized 16 different independent variables, subdivided into 4 conceptual clusters: basic student information (3 variables), basic course/instructor information (4 variables), expected grade in class (1 variable), and student assessments of instructor performance (8 variables). A description of each of these variables is provided below. Basic descriptive data for each indicator can be found in Table 2.

Table 1. Class Information and Total Enrollment for the Data Set Used in this Study

Course Code	Class Description	Total Enrollment Between 2004 and 2009
CJ 101 Justice and Society **(Introduction to Criminal Justice)**	This introduction to the study of crime and justice includes theories and methodologies from a variety of social science disciplines. The course also provides an introduction to the study of social control and to the origins of crime at the individual, structural and cultural levels	4988 (61.5% of sample)
CJ 201 Criminology	An analysis of crime, criminal behavior, punishment, and the theories of deviancy from historical perspectives.	1210 (14.9% of sample)
CJ 300 Research Methods in Criminal Justice	This course involves an examination of basic investigative methods in criminal justice. Focus is on the logic and theory of criminological research, the formulation and testing of hypotheses, research design, sampling, modes of data production, and the ethics of conducting research in criminology and criminal justice.	523 (6.4% of sample)
CJ 312 Police Process	Functions of law enforcement and the roles of the police in contemporary society. Study of the police from several perspectives: historical, sociological, psychological, organizational, and political. Issues, research, and trends pertinent to law enforcement organizations.	643 (7.9% of sample)
CJ 495 Issues in Criminal Justice (Capstone Course)	A capstone course that will entail readings and discussion on contemporary criminal justice issues, ethics, and trends resulting in a senior paper/project.	753 (9.3% of sample)
Total		***n* = 8,117**

Basic Student Information

Three student information variables were used as part of this analysis: student gender, student academic status, and whether the class being evaluated was part of the student's intended major. It was not expected that these variables would have strong influences on perceptions of instructor quality, but it was deemed important to include them as basic controls.

The operationalization of the gender variable was self-explanatory; students were asked to identify if they were male or female. Academic status measured the student's progress

Table 2. Variable Metrics and Bivariate Correlations with the Dependent Variable

Variable	Metrics	Non-Parametric Correlation with D.V.
Student Gender	4098 female, 3848 male	-.011
Student Academic Status	3126 freshman, 1783 sophomore, 1426 junior, 1679 senior, 18 graduate	-.065 **
Class Part of Major	3368 yes, 4604 no	.018
Instructor Gender	1469 female, 6648 male	.054 **
CJ 300 Dummy	523 yes, 7594 no	-.085 **
Strong Desire to Take Course Dummy	5120 yes, 2823 no	.19 ***
General Education Dummy	4741 yes, 3231 no	.058**
Expected Grade in Class	3315 A, 3807 B, 799 C, 45 D, 8 F	.23 ***
Learned a Great Deal in Class	Mean = 3.36	.60 ***
Instructor Challenged Thinking	Mean = 3.21	.55 ***
Instructor Organized	Mean = 3.52	.61 ***
Instructor Made Clear Presentations	Mean = 3.57	.64 ***
Clear Objectives	Mean = 3.48	.61 ***
Responsive to Student Questions	Mean = 3.67	.60 ***
Concern for Student Progress	Mean = 3.34	.57 ***
Fair Evaluation Methods	Mean = 3.38	.61 ***
Overall Instructor Score (D.V.)	Mean = 3.50	-
*** = *p* < .001, ** = *p* < .01, * = *p* < .05		

in the program: respondents were asked to self-identify as freshmen, sophomores, juniors, seniors or graduate students (taking undergraduate courses). The intended major variable was measured as a simple dichotomy: “yes” if the class was in the respondent’s intended major, or “no” otherwise.

Basic Course / Instructor Information

The second cluster of variables assessed basic course and instructor characteristics. Four individual measures were analyzed:

The first was the gender of the instructor: the operationalization of which is self-evident. The second was a dummy variable (“1” yes, “0” no) assessing whether the class in question was CJ 300 (quantitative research methods). The third was another dummy variable measuring whether the student had a strong desire to take the class, and the last was a dummy variable assessing whether the course was being taken as part of the general education program.

Expected Grade in Class

As previously discussed, it is a popular perception that student evaluation scores are less a valid measure of instructor performance, and more a representation of how much students "like" a particular teacher. To this end, some instructors believe that they can become "liked", and thus ensure high SET scores, simply by inflating the grades in their classes To investigate this issue, the present study includes a measure of expected grade as part of the statistical analysis. Students were asked to indicate whether they expected to score an A, B, C, D or F in the class. Unfortunately, the data for this project did not include a measure of the *actual* grade obtained in the class. However, the authors did review data from their own classes and found that the statistical distributions of expected grades, and their actual grade distributions, were relatively consistent. Where discrepancies existed, they tended to be at the lower end of the scale: that is, few students indicated that they expected to receive D's or F's, but a substantial number actually received these grades.

Student Assessments of Instructor Performance

The final eight independent variables are the most theoretically important to this research project. Each of these items represents a dimension of instructor performance. Although there is remarkably little empirical research on the relationships between these types of items and overall perceptions of instructor quality, studies by Feldman (1976), Spencer and Schmelkin (2002), Okpala and Ellis (2005) and Schaefer et al. (2003) do indicate that they may be important. Presumably, instructors who score highly on these 8 items will also be the instructors who obtain the highest overall SET scores.

The individual dimensions of instructor performance are as follows:

> First, students were asked whether they learned a lot from the course in question. Second, they were asked whether the instructor successfully challenged their thinking. Then, they were asked to evaluate whether the instructor was organized, whether he or she made presentations that were understandable, and whether or not he or she conveyed clear objectives for the course. Finally, students were asked whether the instructor was responsive to student questions, whether he or she exhibited concern for student progress, and whether the students perceived the methods of evaluation in the class as fair.

For each item, students were asked whether they strongly agreed, agreed, were neutral, disagreed, or strongly disagreed with the statement under consideration. For the purposes of this project, responses were coded on a 5 point scale ranging from "4" (strongly agree) to "0" (strongly disagree). Although the authors recognize that these variables technically represent ordinal Likert scales, for the purpose of the statistical analysis, they were treated as continuous, interval level measures. This permitted simpler presentation and interpretation of the research results and was not likely to substantially violate the technical assumptions of the analytic techniques used (Glass, Peckham and Sanders, 1972).

It should be noted that the individual indicators described here could have been combined to form a single scale with an alpha reliability of .91. However, because the goal of this project was to determine which indicators had the most powerful influences on overall perceptions of instructor quality, this was not done. Instead, each item was entered into the statistical models individually. Because of the large number of cases available for analysis, this did not create serious statistical issues. All of the models could be successfully estimated, and colinearity remained within generally acceptable boundaries. The highest VIFs appeared in the model containing all 16 independent variables and peaked at 2.69: well below the threshold of 5 proposed by most authors (Kutner, Nachtsheim and Neter, 2004).

Dependent Variable

The dependent variable in this analysis represented a measure of overall student perception of instructor quality. Students were asked to provide a 5 point response (ranging from "strongly agree" to "strongly disagree") to the following question: The quality of the instructor for this course was excellent (try to set aside your feelings for the course). Once again, although this variable technically represents an ordinal level measure, it was treated as an interval scale measure for the purpose of the statistical analyses.

Research Hypothesis

The following section outlines the specific sets of research hypotheses that were tested in this study. The goal of each model is to determine whether the variables analyzed during that step were significantly associated with overall perceptions of instructor effectiveness. The last model, which contains all of the independent variables, provides information for assessing the relative importance of the indicators under consideration.

Model One: Basic Student Information

This regression model tests three null hypotheses. They are as follows: controlling for the other variables in the model, the gender of the student, his or her academic status, and whether the course under consideration is in the student's intended major, will have statistically non-significant effects on that student's perception of the instructor's effectiveness.

Model Two: Basic Course / Instructor Information

This regression model is similar to the first. It tests the following four null hypotheses: controlling for the other variables in the model, the instructor's gender, the student's desire to take the class, his or her taking it to fulfill general education requirements, and whether the course under evaluation is CJ 300 will have statistically non-significant effects on that student's overall perception of the class instructor's effectiveness.

Model Three: Basic Student, Course and Instructor Characteristics + Expected Grade in Course

Model three incorporates all of the variables from models one and two. In addition, this model tests the null hypothesis that controlling for the seven other items in the analysis, the student's perceived grade in the class will have no significant effect on that student's rating of the effectiveness of the course instructor.

Model Four: Basic Student, Course and Instructor Characteristics, Expected Grade, and Individual Indicators of Instructor Performance

Model four incorporates all of the variables from model three. In addition, it tests a series of research hypotheses pertaining to eight individual measures of instructor performance:

1) To the extent that a student perceives that he or she "learned a great deal" from a class, that student's rating of the overall effectiveness of the instructor will be higher.
2) To the extent that a student perceives that the class "challenged his or her thinking", that student's rating of the overall effectiveness of the instructor will be higher.
3) To the extent that a student perceives that the instructor's methods of evaluation for the class were fair, that student's rating of the overall effectiveness of the instructor will be higher.
4) To the extent that a student perceives that the instructor exhibited concern for student progress, that student's rating of the overall effectiveness of the instructor will be higher.
5) To the extent that a student perceives that the instructor was responsive to student questions, that student's rating of the overall effectiveness of the instructor will be higher.
6) To the extent that a student perceives that the instructor made understandable presentations in the class, that student's rating of the overall effectiveness of the instructor will be higher.
7) To the extent that a student perceives that the instructor ran well organized class sessions, that student's rating of the overall effectiveness of the instructor will be higher.
8) To the extent that a student perceives that the instructor clearly communicated course objectives, that student's rating of the overall effectiveness of the instructor will be higher.

Given the logic of multivariate regression analysis, each of the following effects is expected to hold, controlling for all of the other variables in the model.

Analysis

The models were tested using Multivariate, Ordinary Least Squares (OLS) regression analysis (also known as "linear regression"). OLS was chosen for this study because it provides intuitively appealing, and easily interpretable, statistical estimators. This is not always true of other, more complex types of regression techniques. Although the dependent variable in this case is an ordinal scale (as opposed to the interval / ratio level measure typically analyzed using OLS), it was deemed to be a reasonable approximation to a normally distributed measure. This decision is supported by the fact that OLS is widely known to be "robust for validity against nonnormality" (PROPHET Statguide, 1997). Similarly, although mathematically transforming the dependant variable could have normalized its statistical distribution (PROPHET Statguide, 1997), it would have made it impossible to interpret the effects of the independent variables on instructor quality in their original metrics. This would have made the results far less accessible to a general audience, and consequently, the present authors elected not to undertake such transformations.

Results

Table 3, presented below, reveals the results of testing the basic student information hypotheses described above:

Table 3. The Effects of Basic Student Characteristics on Overall Instructor Effectiveness

Variables	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	3.59	.021		170.57	< .001
What is your gender? (Male)	.005	.018	.003	.28	.78
What is your academic status?	-.042	.008	-.062	-4.98	.000
Is this course in your intended major?	.000	.020	.000	-.009	.99
r-squared = .004		F= 9.95 (DF =3)	Sig < .001		

The statistics in this table indicate that the gender of the student, and whether or not the course under evaluation is in the student's major, are *not* significantly associated with perceptions of overall instructor effectiveness. In other words, there was insufficient evidence to reject the null hypotheses. However, academic status *is* significantly linked to instructor effectiveness (p < .001). It seems that more senior students have higher expectations than their junior counterparts. For each step a student advances up the seniority ladder (i.e. from freshman to sophomore, sophomore to junior, junior to senior, senior to graduate student) his or her evaluation of the instructor declines by .04 of a point. Conceptually, this does not represent a huge difference: it suggests that, on average, seniors are approximately .12 of a point "harsher" than freshman. However, the difference has strong statistical significance given the large size of the sample (n =7,796 for this analysis). Table 4 reveals the results of testing the basic course/instructor related hypotheses described above.

Table 4. The Effects of Basic Course and Instructor Characteristics on Overall Instructor Effectiveness

Variables	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	3.14	.027		114.73	< .001
Did you have a strong desire to take this course?	.32	.019	.19	16.74	< .001
Methods Class (CJ 300)	-.025	.038	-.008	-.65	.519
Are you taking this course to fulfill general education requirements?	.14	.019	.083	7.23	< .001
Gender of Instructor (Male)	.091	.023	.044	3.91	.< 001
r-squared = .044		F=89.29 (DF = 4)	Sig. < .001		

The results here indicate the following: Students enrolled in CJ 300 (quantitative research methods) do *not* appear to rate their instructors any more harshly than students enrolled in other courses. However, the other three measures *are* significantly related to overall perceptions of instructor effectiveness ($p < .001$). Students who have a strong desire to take a particular class, those who are taking classes to fulfill general education requirements, and those who have male instructors, are all more likely to rate their instructors positively. Of these effects, the "strong desire to take the class" has the most predictive power (as evidenced by the standardized Beta coefficients). Students who indicated that they had a strong desire to take a class gave their instructors mean evaluation scores that were over .3 of a point higher than students who indicated that they did not have strong desire to take that class.

Table 5 presents the results of the regression model that combined all of the predictors from Tables 3 and 4, and also added a variable measuring the student's perceived grade in the class being evaluated.

This analysis reveals the following: the gender of the student, his or her academic status, whether the class is in the student's intended major, and whether the class under evaluation is CJ 300 have *no* significant impact on the overall instructor effectiveness score (once all of the other variables in the model have been held constant). The reader will recall that these results are consistent with the analyses presented in Tables 3 and 4, except that in Table 3 academic status *was* a significant predictor of perceived instructor effectiveness. Interpretations of this (and other interesting findings) will be presented in the discussion section. Having a strong desire to take the class, taking the class with a male instructor, and taking the class to fulfill general education requirements continue to have significant, and positive, associations with overall instructor effectiveness ($p < .001$ in all cases). These findings are consistent with the results presented earlier.

Table 5. The Effects of Basic Student, Course and Instructor Characteristics, and Perceived Grade on Overall Instructor Effectiveness

Variables	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	2.36	.058		41.04	< .001
What is your gender?	.018	.018	.011	1.01	.31
What is your academic status?	-.007	.009	-.011	-.83	.41
Is this course in your intended major?	-.011	.022	-.007	-.49	.62
Did you have a strong desire to take this course?	.29	.020	.17	14.55	< .001
Methods Class (CJ 300)	.05	.039	.016	1.29	.20
Are you taking this course to fulfill general education requirements?	.14	.023	.087	6.19	< .001
Gender of Instructor	.08	.023	.039	3.54	< .001
Expected Grade	.25	.013	.21	18.60	< .001
r-squared = .086		F=90.90 (DF = 8)	Sig. < .001		

Finally, and perhaps most importantly, the present analysis shows that students who expect to obtain higher grades in a class rate their instructors significantly more generously! For each one grade level change (i.e. from an “F” to a “D” from a “D” to a “C” and so forth) the overall instructor effectiveness score improves by approximately 0.25 of a point. Moreover, the standardized Beta coefficient for this variable (.21) is larger than that for any other variable under consideration. This represents *prima facie* evidence that “easier” instructors are better liked, and tend to obtain higher scores on student evaluations (all else being equal). The differences are conceptually important, students who expect to get an “F” in a class are likely to rate their instructors a full point lower than students who expect to obtain an “A”. However, it is important to note that direct measures of instructor performance have not yet been included in this model. Table 6 presents the results of the regression analysis which did explore this issue.

In the composite model containing all 16 independent variables, the following measures are *not* statistically associated with overall perception of instructor effectiveness: academic status, intended major, strong desire to take the course, the dummy variable for CJ 300 and, most notably, expected grade in course. The fact that one's anticipated grade in the course is no longer significant is a statistically, and conceptually, important difference from Table 5. After the individual instructor performance factors are held constant, students who expect to get high grades in classes no longer rate their teachers any higher than those who expect to receive low grades.

Consistent with the previous models, students in this analysis continue to rate male instructors slightly, but still significantly, higher than female instructors (p = .01) and

Table 6. The Effects of Basic Student, Course and Instructor Characteristics, Perceived Grade, and Individual Instructor Performance, on Overall Instructor Effectiveness

Variables	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.28	.047		-6.03	< .001
What is your gender?	.02	.011	.014	1.97	.049
What is your academic status?	.001	.006	.002	.21	.83
Is this course in your intended major?	-.009	.014	-.005	-.62	.54
Did you have a strong desire to take this course?	-.02	.013	-.012	-1.54	.12
Methods Class (CJ 300)	.04	.024	.012	1.64	.10
Are you taking this course to fulfill general education requirements?	.034	.014	.021	2.38	.017
Gender of Instructor	.036	.014	.017	2.48	.013
Expected Grade	-.013	.009	-.011	-1.47	.14
Learned Great Deal in Class	.20	.01	.20	19.24	< .001
Material Challenged Thinking	.11	.009	.12	12.94	< .001
Methods of Evaluation Fair	.14	.01	.14	13.94	< .001
Instructor Concerned Student Progress	.11	.009	.11	11.65	< .001
Instructor Responsive to Questions	.17	.013	.13	13.13	< .001
Instructor Presentation Understandable	.20	.013	.17	14.90	< .001
Instructor Organized Class Sessions	.10	.013	.09	7.69	< .001
Course Objectives Clearly Communicated	.08	.012	.07	6.32	< .001
r-squared = .65		F=859.15 (DF = 16)	Sig. < .001		

classes taken as part of the general education program somewhat higher than those not taken to satisfy general education requirements (*p*=.02). However, contrary to previous models, the sex of the student now significantly impacts perceptions of instructor performance (with males giving marginally higher scores, *p* = .05), while having a strong desire to take a course is no longer a statistically significant predictor (the reader should recall that this was the *strongest* significant predictor in Table 4).

The most striking findings from the present analysis; however, pertain to the eight individual instructor performance variables. All of these variables are strongly, and positively associated with the overall instructor evaluation score (even after controlling for the other variables in the model, *p* < .001 in all cases). Students who agreed that they learned a great deal in class, that the class challenged their thinking, that the instructor's methods of evaluation were fair, that the instructor demonstrated concern for their progress, that he or she was responsive to student questions, that he or she made understandable presentations, that class sessions were well organized, and that course objectives were clearly communicated, consistently rated their instructors higher than those who disagreed with these statements. The two strongest items (as measured by the Beta coefficients) turned out to be "learned a great deal in class" and "instructor made understandable presentations". For each extra point on these two items (i.e. moving from "agree" to "strongly agree" and so forth) students assigned their instructors approximately .2 of an extra point on their overall evaluations.

Discussion

The most important findings from the regression analysis are that although student/course/ instructor characteristics like gender, nature of the class under consideration, etc. can influence overall perceptions of instructor effectiveness, their overall impact is small: particularly when compared to more direct measures of instructor performance. Students who expect to score high grades in classes *are* more likely to rate their instructors higher in the statistical model that controls only for basic characteristics (Table 5); however, they are *not* any more likely to favor their instructors after measures of instructor performance have been held constant (Table 6). This suggests that what an instructor chooses to do in class is more important to his or her overall SET scores than more superficial variables. More specifically, this study provides statistical evidence to support the idea that well prepared, organized, caring, responsive, fair instructors will be rated higher than those who struggle in these areas. These finding are consistent with the research conducted by Crumbley, Henry, and Kratchman (2001) and helps to validate subjective instructor evaluations as a legitimate measure of teaching effectiveness (See also Spencer & Schmelkin, 2002; Okpala & Ellis, 2005; and Schaefer et al., 2003). Our study also helps dispel myths like: teachers can "buy" high evaluation scores with high grades, or teachers who are assigned "difficult" classes like quantitative research methods are doomed to receive low evaluation scores.

Given the variation between these findings and others, research in this area should seek to establish a reliable estimate of the strength of the relationship between grades and overall SETs. It is possible that high grades are actually a result of effective teaching, as, students who learn more tend to give their teachers higher overall evaluations (Marsh & Roche 1997). However, to the extent grades significantly effect student perceptions of effective teaching, questions about the value of SETs will remain.

The finding that male instructors seem to be rated consistently higher than female instructors is also conceptually interesting. Perhaps it reflects some inherent bias against women teaching at post secondary institutions. However, the present researchers would like to offer a word of caution as these findings are based on evaluations of faculty in a school of criminal justice. As such, these findings may be less applicable in other disciplines given that criminal justice is traditionally a male dominated discipline. It may be possible that students do have a gender bias when the subject is criminal justice: much like Gabbidon (2002) and Hase (2001) found with classes pertaining to race and culture taught by minority professors.

Based on the present analysis, it is possible to assert that this gender effect does *not* reflect that male instructors simply perform better in the classroom than females. Were this true, one would have expected this effect to statistically "disappear" once direct measures of instructor performance were controlled in Table 6: yet this did not occur. Unfortunately, this study does not provide any additional data to examine possible explanations for this phenomenon. Moreover, because our sample was drawn from only one Midwestern university, there is no guarantee that even this basic finding will generalize to other institutions. As noted earlier, the literature has provided inconsistent results in this area. Consequently, it seems most appropriate at present to merely note this finding as interesting

Conclusions and Limitations of the Present Study

Overall, this study provides preliminary evidence to support the notion that SET evaluations may be a valid, and reliable, method of assessing instructor performance. The analysis suggests that faculty who care about their students, are well organized, perceived as fair, and who inspire student thinking, seem to be perceived as the most effective classroom teachers. These instructor performance items cluster quite reliably (alpha = .91), and are more strongly connected to overall perceptions of teacher quality than descriptive student, course, or instructor characteristics. They also seem to matter more than the ultimate grades that students expect to receive in their classes.

When taken together, these findings provide an encouraging message: they suggest that students want, and expect, far more out of their classes than simply "getting an A". They want instructors who value them, feel passionate about the subjects that they teach, and are able to inspire student learning. As college tuition continues to get more expensive, and employers continue to become more demanding of college graduates, it seems plausible that these trends will continue. When one is paying thousands of dollars for an education, it seems only reasonable that one would expect to walk away with more than "a piece of paper" provided by an indifferent faculty and/or institution.

That said, the authors do not wish to claim too much for the present analysis. The speculative conclusions offered in the previous paragraph assume that one takes the findings at face value. There are a number of methodological issues with the present research that warrant caution (particularly when applying the results to other post-secondary institutions).

First and foremost, the reader should recall that this study is based on a non-probability sample of criminal justice students enrolled in a single SCJ, at a single Midwestern university. To the extent that other institutions, and their student populations, are substantially different, these findings may not generalize. More specifically, the present authors would recommend caution when applying the findings to prominent tier one research universities, community colleges, or small liberal arts colleges (i.e. places that are substantially different from the university under study). Administrators at institutions where most classes are taught by part time faculty (i.e. doctoral students, adjuncts etc.) should also be cautious about assuming that these findings will apply to their schools, as should those where much of the student body consists of foreign, or minority students (as this university draws primarily from one region in one state, and has a student body that is disproportionately White and middle class.

It should also be noted that the present findings are cross-sectional. Therefore, readers should be cautious in assuming that any of the associations revealed here represent true, causal, relationships. For example, it is the present author's assumption that the "grade effect" revealed in Table 5 is spurious to the individual instructor performance items analyzed in Table 6. That is, that instructors who grade fairly, inspire their students, make interesting presentations etc. will inspire their students to earn higher grades. However, it is also plausible that students who have earned high grades will be retrospectively more likely to rate their instructors as fair, inspiring, interesting etc. Because the variables in this study are measured at one point in time, the present researchers have no way of determining which causal sequence is more accurate.

Finally, it is clear that some of the measures used in this study are not ideal. Most notably, to be truly confident about "the grade effect" it would have been useful to have a measure of *actual* grades earned in the class, as opposed to *perceived* grades. As noted earlier, the authors' informal analysis suggested there is a fair bit of overlap in these measures within the authors' own classes; however, this does not necessarily have to be the case with all instructors. It is not implausible that some instructors give students the impression that they will receive grades that are substantially lower (or higher) than the actual grades that appear on their transcripts. Further validating this, and the other measures in this paper, would allow researchers to have a higher degree of confidence in the findings of this research.

Despite these limitations, it does not appear that any of these shortcomings are sufficiently serious as to render the results of this paper meaningless. On the contrary, the present authors would like to suggest that this study makes an important contribution to the debate on the validity of SETs as measures of instructor performance. However, the researchers suggest further inquiry in this area to shed more light on the issue of how student evaluations can best be used as formative, developmental, and summative evaluation tools, within the discipline of criminal justice.

References

Centra, J. A. (1979). *Determining Faculty Effectiveness*. San Francisco: Jossey-Bass.

Chako, T, L. (1983). Student Ratings of Instruction: a function of grading standards. *Educational Research Quarterly*, 8, 19-25.

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

Crumbley, L., Henry, B. K., & Kratchman, S. H. (2001). Students' perceptions of the evaluation of college teaching. *Quality Assurance in Education*, 9, 197-207.

Eison, J. (1999). Commonly Asked Questions about Student's Ratings of Courses and Instructors. Center for Teaching Enhancement: University of South Florida.

Ellis, R. S. (1985) Ratings of Teachers by their Students should be used Wisely-- Or not at all. *The Chronicle of Higher Education* 31:88.

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

Franklin, J. L., & Theal, M. (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*, 109: 45-56.

Gabbidon, S. L. (2002). Exploring the Role of Race and Course Content in Criminal Justice Course Evaluations: A Case Study. *Journal of Criminal Justice Education*, 13, 1, 101-112.

Glass, Peckham, & Sanders (1972). Consequences of Failure to Meet Assumptions Underlying the Analyses of Variance and Covariance. *Review of Educational Research,* 42, 237-288.

Gleason, M. (1986). Getting a Perspective on Student Evaluation. *AAHE Bulletin.* February, pp. 10-13.

Guthrie, E. R. (1954). The Evaluation of Teaching: A Progress Report. Seattle: University of Washington.

Griffin B. W., & Pool, H., (1998). Monitoring and Improving Instructional Practices. *Journal of Research and Development in Education,* 32: 1-7.

Hase M. (2001). Student Resistance and Nationalism in the Classroom: Some Reflections on Globalizing the Curriculum. *Feminist Teacher* 13:90-107.

Hendrix, K. G. (1998). Student Perceptions of the Influence of Race on Professor Credibility. *Journal of Black Studies*. 28:738-763

Howard, G., Conway, G., & Maxwell, E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Education Psychology*, 77:187-196.

Kulik, J. (2001). Student Ratings: Validity, Utility, and Controversy. *New Directions for Institutional Research*, 109:9-25.

Kutner, M., Nachtsheim, C. & Neter, J. (2004). Applied Linear Regression Models, 4th edition, McGraw-Hill Irwin.

Lersch, K., & Greek, C. (2001). Exploring the Beliefs Surrounding Student Evaluations of Instruction in Criminology and Criminal Justice Undergraduate Courses. *Journal of Criminal Justice Education,* 12, 2, 283-299).

Lueck T. L. (1993). The interaction effects of gender on teaching evaluations. Journalism Educator, 48:46-54

Mangan, K. (2009). Professors Compete for Bonuses Based on Student Evaluations. *The Chronicle of Higher Education*, 55, 21, A10.

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

Marsh, H. W., & Dunkin M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher Education: Handbook on Theory and Research* (Vol. 8, pp. 143-234). New York: Agathon Press.

Marsh, H. W., & Roche, L. A. (1997) Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility. *American Psychologist*, 52, 11, 1187-1197.

Miron, M., & Segal, E. (1986). Student Opinion of the Value of Student Evaluations. *Higher Education*, 15, 3, 259-265.

Murray, H. G. (1980). Evaluating university teaching: A review of research. Toronto, Ontario, Canada: Ontario Confederation of University Faculty Associations.

Neath, I. (1996). How to Improve your Teaching Evaluations Without Improving Your Teaching. *Psychological Reports*. 78:1363-1372

Okpala, C. O., & Ellis,R. (2005) The perceptions of college students on teacher quality: A focus on teacher qualifications. Education, 126, 374-378.

PROHPHET Statguide. (1997). Online resource provided by Northwestern University, Evanston: IL.

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

Rowden, G. and Carlson, R. (1996). “Gender Issues and Students’ Perceptions of Instructors’ Immediacy and Evaluation of Teaching and Course.” *Psychological Reports* 78: 835-839.

Ryan, J., Anderson, J., & Birchler, A. (1980). Student evaluation: The faculty responds. *Higher Education* 12, 4, 395-401.

Sandler, B. R. (1991). Women Faculty at Work in the Classroom, or Why it Still Hurts to be a Woman in Labor. *Communication Education.* 40: 6-15.

Schaeffer, G., Epting, K. , Zinn, T., & Buskit W. (2003). Student and faculty perceptions of effective teaching: A successful replication. Teaching of Psychology 30, 133-136.

Spencer, K. J., & Schmelkin, L. P. (2002). Students' perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education*, 27, 397-408.

Seldin, P. (1993). How Colleges Evaluate Professors: 1983 Versus 1993. *AAHE Bulletin*, Oct. 1993, pp. 6-8, 12.

Scriven, M. (1983). Summative Teacher Evaluation. In J. Milman (ed.), *Handbook of Teacher Evaluation*. Thousand Oaks, CA: Sage.

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(2), 355-376.

Theal, 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*, 109, 45-56.