INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

AN EXAMINATION OF INSTRUCTIONAL EFFECTIVENESS IN HIGHER EDUCATION USING MULTIPLE OUTCOME MEASURES

BY

MARK LAWRENCE WILSON

A DISSERTATION PRESENTED TO THE

GRADUATE FACULTY OF MIDDLE TENNESSEE STATE UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF ARTS

MAY, 1998

AN EXAMINATION OF INSTRUCTIONAL EFFECTIVENESS IN HIGHER EDUCATION USING MULTIPLE OUTCOME MEASURES

APPROVED:

Major Professor

Committee Member

Committee Member

Chairman of the Department of Economics and Finance

Dean of the College of Graduate Studies

ABSTRACT

An Examination of Instructional Effectiveness in Higher Education Using Multiple Outcome Measures

by Mark Lawrence Wilson

This study develops a technique of evaluating teaching effectiveness in higher education using two outcomes measures. The first measure, an instructor's score on the student evaluation of the teacher (SET), is considered to be the traditional measure of teaching effectiveness. The second measure, a comprehensive exam, estimates the cognitive performance of each teacher's students. These measures are merged to create a *composite* index, and this index is used to measure teaching effectiveness.

It is maintained that accurate performance measurement facilitates labor allocation and motivation schemes. Traditional faculty performance evaluation may elicit unfavorable economic consequences due to perverse incentive effects on teaching behavior such as *moral hazard* and grade inflation. A broader measure of teaching effectiveness may lessen some of the perverse incentives of the traditional system.

The first step to create the composite index is the estimation of educational production functions for the SET and the outcomes exam. The estimation uses both panel data and Seemingly Unrelated Regressions (SUR) models. Second, a predicted score is estimated for all instructors using the sample average values of the right-hand side variables. Third, the predicted score is compared with the actual score to identify

exceptional performers. Finally, the measures are weighted to develop a composite score for each instructor.

The data for this study were collected from a survey administered to students, a comprehensive exam, and administrative records. The data were collected at a large, comprehensive university in the Fall semester of 1996. The population of approximately 700 students came from 24 classes of introductory economics taught by twelve instructors.

Under the panel data technique, it was found that student evaluations of the teacher are influenced by a student's expected grade, choice of section, native language of the instructor, academic major, and student aptitude. Cognitive learning is influenced by student aptitude, grade point average, age, and male gender. Using the SUR estimation method, several variables were added to each learning equation. For the SET estimation, the large class variable had a negative coefficient, whereas the instructor's years of service has a positive coefficient. For the exam equation, high school economics was significant with a negative coefficient and the instructor's terminal degree entered the equation with a positive coefficient, but the male and age variables were insignificant.

The composite index re-orders the measurement of faculty performance. It was observed that some outstanding performers were overlooked by the traditional measurement technique. Some exceptional performers by the traditional technique were deemed to be average performers by the adjusted measure. The study also found evidence of grade inflation.

ACKNOWLEDGMENTS

A study of this size could not be completed without the help of many people. I would like to thank several of those who have contributed greatly.

First, I would like to thank my family. My wife, Margaret Cameron Wilson, deserves great credit for the final product. She was very loving and supportive, and also managed the household with great control during these years when there were four students in the house. My children Alexander, Peter, and Francis also deserve many thanks for being fun-loving kids when their mother and father were sometimes distracted. I also am indebted to my parents. My mother, Marilyn I. Wilson, deserves many thanks for the encouragement and love she has given me over the years. Also, I would like to thank my father, the late George "Link" Wilson (d. 1977). Although he did not live to share my graduate education, his early influence and guidance made it possible.

I would also like to thank the three members of my outstanding dissertation committee. First, I am grateful to the supervisor, Professor Joachim Zietz, for his original ideas, quiet encouragement, and careful reading. Professor Reuben Kyle also made substantial contributions to this study, and was especially helpful in keeping me aware of the underlying economic theory. Professor James Huffman is much appreciated for his kind comments and for giving me a chance to present this study in his classroom.

Finally, I would like to thank Professor John Lee, Chairman of the Economics and Finance Department, for helping me secure financial aid. Also, thanks to Professor Billy Balch for his role as my academic advisor, and to the graduate faculty of the Economics and Finance Department for their outstanding instruction and high scholarship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

Appendix **Page**

CHAPTER 1

INTRODUCTION

This study examines the measurement of teaching effectiveness in higher education. The term "teaching effectiveness" is a complex mix of faculty and student inputs that produce educational outputs. Because both faculty and students are involved in this process, and because both have different personalities, interests, and abilities, the measurement of teaching effectiveness is not reducible to a single value. Instead, a broader approach is desirable. The approach of this study is to collect two measures of teaching output, adjust these measures for any extraneous influences, and use these two measures to form a composite teaching effectiveness score. In this way, the measured value of teaching effectiveness will incorporate, to the extent possible, the many inputs and outputs of higher education.

Examination of educational research reveals that many articles and studies have been published about teaching effectiveness. But most of these studies, particularly in the field of economics, have only identified *one* measure of teaching effectiveness. It is the thesis of this research that, given the complexity of student-faculty interaction, more than one student output must be considered. The relative importance of these multiple outputs can then be determined by administrators.

Reliable measurement of teaching effectiveness should be of value to all participants in the higher education setting: the administration, the faculty and the student body. Administrative decisions are improved when teaching effectiveness is accurately measured

because faculty assignments and reward patterns are more easily established. Similarly, faculty members are influenced by the measurement of teaching effectiveness insofar as their work assignments, incentives, and professional rewards are affected. Finally, students are beneficiaries of accurate measurement of teaching effectiveness insofar as this influences the hiring, retention, and allocation of the instructors they face in the classroom, in addition to the quality of the students' education.

1. Statement of the Problem

Colleges, universities and other institutions of higher education produce many services. The most well known of these services are research and teaching.[1] In a purely economic context, one may be concerned with how these educational outputs are measured. Adequate measurement of outputs is essential for cost and benefit comparisons. As for research outputs, the academic community shares widespread agreement about a faculty member's publication record. The number of publications can be counted and the prestige of the journals quantified.[2] As for teaching, however, the standard by which output is measured is much less clear than the publication standard. Teaching produces several educational outputs, and these outputs often overlap. Many of the outputs have both private and social benefits. For the sake of organization, the outputs considered in this study are termed *cognitive* and *affective* outputs. Cognitive outputs are content-area knowledge outputs, such as the ability to compute the *investment*

[1] Other services are produced as well. For example, Armen Alchian says, "It is my general premise that the university has two functions; I'm not sure which is the most important one. One is education, the other is a marriage market." (Alchian et al. 1996, p. 426)

[2] An example of ranked economics journals can be found in an article by Gibbons and Fish (1991, p. 363).

multiplier of government expenditures in economics. Affective outputs are attitude-related outputs, and are associated with a student's taste or distaste for the subject matter, their taste for the instructor, their interest in reading about the subject in the future, and so forth.[3]

Having made the distinction between educational outputs, it is necessary to develop some ideas about the economic production of these outputs, and the administrative and faculty behavioral models that underlie this process. Commonly, the production of these outputs is modeled like many other production processes; namely, inputs are transformed into finished outputs. This input-output relationship, known as a *production function*, assumes that a transformation process is going on. That is, given some fixed level of technology, inputs are being improved upon and transformed into more valued outputs (Hanushek 1979). According to this reasoning, an *effective* teacher displays a higher than average rate of transformation. It is this rate of transformation, or teacher effectiveness, that is measured by this study.[4]

Assuming that teacher effectiveness can be accurately measured, both college administrators and faculties value this information. First, this topic is considered from the perspective of the administration.

Administrators, in their economic roles, are concerned with resource allocation and the reward structure of their employees. In terms of resource allocation, administrators are motivated to deploy their faculty in the most efficient way possible. This means that

---

[3] Thanks to Dr. Duane Graddy, Middle Tennessee State University, for making this distinction. It is also made by Williams and Ware (1976), Lima (1981), and Saunders (1990).

[4] Faculty research output, of lesser importance in this study, is considered later as an administrative issue.

administrators make decisions to allocate faculty resources among research and teaching assignments in a way that will generate the most educational output for the fixed academic budget.[5] These decisions require at least two preconditions: the ability to move resources between teaching and research inputs, and the ability to measure the research and teaching outputs. Assuming that a faculty member's time can be moved among teaching and research outputs, and assuming that research outputs can be readily measured, the accurate knowledge of teaching output and effectiveness is required for optimal resource allocation.

Administrative reward patterns, especially promotion and remuneration, are also influenced by accurate measurement of teaching effectiveness. It is well known that most colleges and universities use measures of teaching effectiveness in hiring, promotion, tenure, and merit pay decisions (Costin, Greenough and Menges 1971; Dilts and Fatemi 1982). For that reason, administrators may value accurate and reliable measures of teaching effectiveness. Absent a reliable measure of teaching effectiveness, administrators may rely on biased or incomplete information about a faculty member's classroom performance.

From the faculty perspective, the measurement of teaching effectiveness is also valued. Because measured teaching effectiveness influences resource allocation and rewards in the present, this measure also influences behavior and incentives in the present and the direction of a teacher's career in the future. With respect to incentives in the

---

[5] Faculty members also produce *service* outputs, which consist of consulting activities, committee work and advising. Because service output seems to be nearly uniform among faculty, and because service is a smaller fraction of output than research and teaching, it is disregarded in this study. It should be noted that service outputs appear to be increasing in importance over time.

current period, and absent accurate measurement of teaching effectiveness, faculty have an incentive to exaggerate their current effectiveness and take least-cost means of improving measures of effectiveness. Least-cost methods of improving measures of effectiveness may include easier grading patterns or content debasement. In the future, the *expansion path*, or the allocation of faculty outputs in the next period, is influenced by the resource allocation pattern in the current period. All told, the measurement of teaching effectiveness has significant reward and incentive implications for the faculty.

To summarize, the adequate measurement of teaching effectiveness has significant educational and economic implications. These implications are seen in the allocation and reward structure of administrators, the behavior, incentive and career paths of the faculty, and the learning outcomes of students.

2. Objective of the Study

The objective of this study is to measure teaching effectiveness in the college classroom and use this measurement to rank instructors. While measuring teaching effectiveness is a regular practice throughout colleges and universities, most measurement involves a nebulous definition of what is being measured. That is, the traditional measure of teaching effectiveness is the collection of student evaluations of the teacher (SETs). These aggregate SETs provide the basis for inter-faculty comparisons. While this technique has the advantage of cheap and quick data collection and interpretation, this method suffers from a flaw: the students' learning is not directly measured. If the SET does *not* measure cognitive learning, and if cognitive learning is valued by the educational institution, then measures other than or in addition to the SET must be used. Because the

evidence is strong that SET scores are not coextensive with cognitive measures (Needham 1978; Williams and Ware 1976), some attempt to measure broader educational outcomes is appropriate. It is the objective of this research to collect multiple educational outputs, merge these outputs into a composite teaching effectiveness score, and rank the faculty on this composite score. This composite score is then compared with the unadjusted score and tested for the consistency of the two measures.

Realization of this objective is complicated by the fact that educational outcomes are most likely related to influences other than the teacher. That is, institutional factors such as the time-of-day the class is taught and the size of the class, as well as other factors, may influence student performance. Similarly, the composition of class varies from section-to-section. Students vary in the amount of their ability, motivation, background, among other relevant characteristics, and these differences may contribute to their evaluation of the teacher and learning rate. Therefore, these potential influences must be controlled so that extraneous forces do not influence the measure of teaching effectiveness. After adequate control variables are identified, an instructor's gross cognitive and affective scores are *mean-leveled.* Simply stated, this mean-leveling process creates a "what if" scenario. That is, *if* an instructor got an *average* class, in all respects, what would be the instructor's teaching effectiveness score? Conceptually, this mean-leveling extracts all of the factors that are outside the instructor's control, and thus provides the basis for inter-instructor comparison of teaching effectiveness.

3. Uniqueness of the Study

This research makes unique contributions to both the theoretical and empirical studies in the economics education research. The following discussion elaborates on these contributions.

The main theoretical contributions of this research are twofold. First, this study advances the idea that the student has volition, or choice, in the selection of their instructor. Moreover, this volition should be a formal part of studies that evaluate a student's classroom performance because the student may consider teachers' reputations for difficulty, humor, and other attributes. The extant literature in this area of economics education generally disregards student selection of teacher and section.[6] This amounts to an implicit assumption that students are randomly assigned to their instructor. This paper makes explicit recognition that students *choose* their class setting, although this choice is constrained by class availability.

The second theoretical contribution of this research is the creation of a *composite* instructor-ranking scheme. It is well known that teaching effectiveness has been evaluated and ranked by several criteria. Usually, however, the ranking schemes are based on the measurement of only one faculty attribute, such as SETs, peer evaluations, or unsolicited student comments. The ranking scheme designed in this paper measures two classroom outputs, cognitive and affective learning, and uses these outputs to form a composite measure of teaching effectiveness. The advantages of a composite measure are several. First, because both of these outputs seem to be desirable educational outcomes,

[6] Two exceptions to this are the articles by Leventhal et al. (1975), and Wetzstein, Broder, and Wilson (1984).

it is useful to measure both. Also, measuring more than one educational output allows for flexibility in the interpretation of teaching effectiveness. That is, if either of these outputs is more highly valued by a school, then the institution can weight the categories to reflect the school's individual mission and priorities. For example, a prestigious liberal arts college that has a long-standing reputation for placement of economics majors into graduate schools may choose to emphasize the cognitive output of their students. Conversely, a college that emphasizes faculty-student interaction may well choose to identify teachers that excel in the production of affective educational outputs.

This research also makes a unique empirical contribution to the economics education literature in the use of a large, individualized data set. Although other large data sets have been used in this area of economics education, the preponderance of the studies have used *class average* SETs and achievement scores, and compared these with instructor behaviors and control variables. This method of using section-by-section averages is called the "class" method by Kau and Rubin (1976), and is distinguished from the "student" method used in this study. The latter method uses individuals as data points instead of class averages. Student level data are more difficult to collect, because each student must individually agree to be identified. If collected, however, the statistical precision of this method is greater because there is more variation among several hundred students than there is among a much smaller number of classes. In statistical parlance, this variation provides more *degrees of freedom*, and allows for the more precise fitting of statistical relationships.

4. Limitations to the Study

The data set collected for this study is sufficiently large to enable robust statistical inferences to be made and the economic theory presented is consistent with the neoclassical utility theory and the theory of the firm. Nevertheless, some limitations of this research must be proffered.

There are four apparent limitations to this study: the use of self-reported survey data for some of the student-specific variables; the limitations of the sample due to withdrawals from the class and absences on the day the SET was administered; diverse student preparation and incentives for the achievement exam; and the absence of a generally agreed upon theory of cognitive learning. Each of these will be addressed in turn.

Self-reported data, although widely used in economics education research, may suffer from a lack of accuracy (Maxwell and Lopus 1994; Valenzuela and Dornbusch 1994). Some self-reported inaccuracies are simply due to unclear survey questions, while other inaccuracies are due to forgetfulness and uncertainty. Sometimes self-reported errors are intentionally made to subvert the purpose of the research. In any event, cross-checks for accuracy are required, to the extent possible, whenever self-reported data are used as the basis for making statistical inferences.

Second, economics education research is complicated by students who withdraw from classes and from student absences on the days that survey information is collected. In this study, these potential problems are recognized and remedial steps are taken to account for their effects.

Third, the achievement exam that was written and administered to this sample was taken under non-uniform conditions. That is, because the exam was not a departmental requirement, the incentive to perform may have varied from class-to-class. Some instructors used the exam as a comprehensive final with a large influence on the students' course grade. Other instructors did not prepare their classes for a comprehensive final exam and, for that reason, the exam was minimally weighted. Some instructors did not use the exam at all. Moreover, since the exam was written by different instructors from those who taught the classes, there may be inconsistencies in the weight of the topics covered and the testing style. These effects were mitigated, to some extent, because all twenty-four classes used a common textbook and the comprehensive exam was designed using many questions from the test bank that accompanied the textbook. Nevertheless, these limitations must be explicitly recognized.

Fourth, research in economics education has been hobbled by the lack of a standard theory of cognitive learning (Saunders 1990). Simply stated, it is not entirely clear how students are motivated, what instructional techniques are superior, or which characteristics the students ultimately value in an instructor. These uncertainties have been the cause of much conflicting evidence in this area of research.

5. Organization of the Study

This research is organized in six chapters, and follows the general pattern that is detailed below.

Chapter 2 is a review of the relevant literature. This literature review includes a discussion of the economic models that are used in this study, a discussion of the process

and empirical results of SETs, the process and empirical results of achievement exams, and finally, a review of those studies that have attempted to rank teaching effectiveness in a non-aggregate way.[7]

Chapter 3 discusses the economic models that are impinged upon by the measurement of teaching effectiveness. First discussed is a model of administrative allocation and reward schemes. This model contemplates the maximization of the academic outputs of research and teaching bounded by a resource constraint. Also discussed is a model of faculty behavior based upon their constrained utility maximization. These two models form the theoretical backbone of why accurate measurement of teaching effectiveness is important.

Chapter 4 discusses the empirical estimation of the teaching effectiveness scheme that is developed in this research. The data are evaluated in terms of their source and quality, and statistical methods of estimation are discussed. The section concludes with a discussion of the criteria for the appropriate estimator selection, and explains the choice that is made.

Chapter 5 reports the empirical findings of this research. This includes the final variable selection, coefficient estimation, and measures of statistical adequacy for both of the educational production functions that are estimated. Also included is a discussion of the conformity of the results with the economic model and a possible explanation of any unexpected findings. The chapter concludes with a comparison of the rankings of faculty

[7] *Aggregate* SETs are defined as raw values of the student evaluation of the teacher. *Adjusted* SETs are defined as those student evaluations that have been adjusted, in some way, for extraneous factors.

on an aggregate basis versus the adjusted basis suggested by this paper. A statistical comparison of these rankings is also made.

Chapter 6 provides a summary of the research and explains the totality of the findings of this study. The chapter concludes by explaining the potential application of the proposed technique, and makes suggestions for further research. The conclusions of the research are also stated.

CHAPTER 2

REVIEW OF THE LITERATURE

The studies in economics education that have examined the issues addressed in this paper can be identified as the "faculty performance measurement" literature. To organize the contributions to this literature, it is helpful to identify several relevant themes. There are four themes of particular importance: 1) theoretical models of faculty performance and administrative behavior; 2) measurement of teaching effectiveness using SETs; 3) measurement of teaching effectiveness using achievement tests; and, 4) non-traditional faculty ranking schemes.

The economic models in this literature address the incentives and behavioral tendencies of the economic agents that are involved in higher education. These models use the core of labor economics and utility theory to derive predictions about the behavior of these agents. In this research, the economic agents of interest are the administrators and faculties of colleges and universities.

The second theme of interest is the measurement of faculty performance. This is empirical research, and has usually involved the application of an educational production function to the classroom. This production function statistically estimates the relationships between student outcomes and educational inputs, with special consideration of the instructor's inputs.

The final theme to be discussed is the faculty measurement and ranking schemes that attempt to distinguish the effect of the teacher from the non-teacher influences of the

class. These ranking schemes take an aggregate measure of teaching output and adjust this measure for student and institutional influences. The result is a "net" teacher score, which is then used to rank instructors. The theoretical contributions to this literature, the empirical studies and the ranking schemes are enumerated below.

1. Models of Faculty Performance and Administrative Behavior

Economic models that address the issue of faculty performance measurement can be broken down into issues that affect administrators, faculty, and students.[14] Administrative issues address the resource allocation function and reward schemes of higher education. Faculty issues are captured by labor supply models and their attendant reward schemes and incentives. Student issues are captured by models of production functions applied to higher education. Each of these three major groups will be discussed in turn.

1.1. Models of Faculty Resource Allocation and Reward Structure

This strain of the literature, largely the work of Josef Broder and William Taylor (1994), deals with the faculty-administration economic relationships that are found in colleges and universities. In general terms, administrators are charged with much the same mandate as the typical manager of a business firm: maximize output subject to a resource constraint.[15] The faculty members are the economic agents who are the subjects of this allocative process. Their faculty output consists of teaching "units" and research "units,"

---

[14] This research is primarily interested in the economic models that apply to the administration and faculty. Nevertheless, the student models are included here because they cannot easily be extricated from the faculty models.

[15] Similarly, minimization of costs subject to a fixed level of output will generate the same outcome if it is assumed that colleges and universities behave like business firms.

and the allocative process involves moving labor resources between these outputs to attain the optimum output mix.

In the bulk of neoclassical economic theory, the price system serves to allocate resources. But because there are no clear market prices for the output of the faculty, other implicit methods of pricing must be relied on (Broder and Taylor 1994).[16] As discussed earlier, the reputations of academic journals and their referees serve an important economic function as a gauge of research quality. Journal publications implicitly measure the research output of a faculty member.[17] For teaching, however, no clear external standards exist. For this reason, some internal standards must be created. One solution is that faculty members are partially responsible for providing information about their effectiveness through promotion dossiers and portfolios. SETs, peer evaluations, and the evaluations of administrators provide the remaining information on which administrative decisions are made.

Broder and Taylor developed a theoretical model that explains the reasons why accurate measurement of faculty outputs may have significant economic effects. They base their arguments on the *imperfect information* problem of microeconomics. That is, most models of microeconomic behavior are based on situations where both parties to an exchange have perfect information. If this condition does not hold, then there is a situation of *imperfect information*, and economic models must be adjusted accordingly.

---

[16] One might argue that relative college tuition rates are the market prices of schools, and thus reflect the value of a faculty member's output. While this might generally be true for the school as a whole, it is less applicable when evaluating an academic department of that school.

[17] Participation in academic conferences and the writing of in-house working papers are important as well, and some of this output could be overlooked by a "citation-count" standard. It should be noted that these works often result in publications.

Occasionally there is imperfect information where one party has better information than the other party. This is the case of *asymmetric* information. Clearly, the economic agent with the better information is in a position to economically exploit the other party. If the asymmetry is not extreme, or if the economic stakes are low, a market characterized by these conditions can continue to function. However, if the information asymmetry becomes too extreme, or the economic stakes are too high, the market breaks down and trade stops. This is a case of the "Lemons Problem" that was identified by George Akerlof (1970).[18]

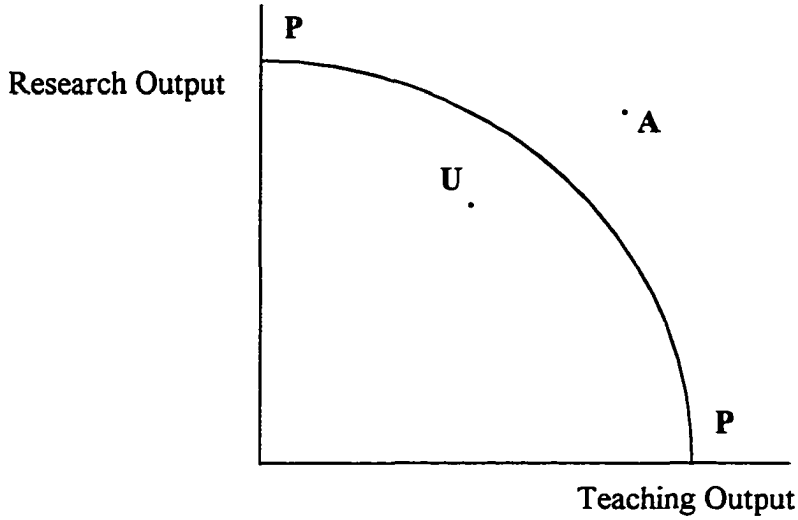
In the case of higher education, the students and the instructor have a great deal of knowledge about what happens in the classroom but the administrators do not. Absent any student or peer evaluations of the teacher, the administration would rely on the instructor-reported measures of teaching output and unsolicited comments by students and faculty. Clearly, this asymmetric information situation gives a clear advantage to the instructor. Faculty performance measurement then, allows the administration to acquire information from *all* the participants in the classroom setting, and thus mitigate the problem of asymmetric information. According to Broder and Taylor, this measurement allows the construction of an accurate production possibilities frontier for the academic outputs of teaching and research. This is labeled PP in Figure 1

One additional clarification is needed. In Figure 1, the outputs of the faculty are labeled "Research" and "Teaching." The research axis has been adequately explained, but

[18] Akerlof (1970) described the breakdown of exchange when the "lemons problem" occurs. Briefly, if a seller has good information about the quality of a product, and knows the product is of high quality, but the prospective buyer has little information about the product, and assumes the seller is selling an inferior product (a "lemon"), then the exchange is not consummated, and trade breaks down. Exchange, in this case, will resume only when information improves.

Figure 1.

Faculty Production Possibilities Frontier

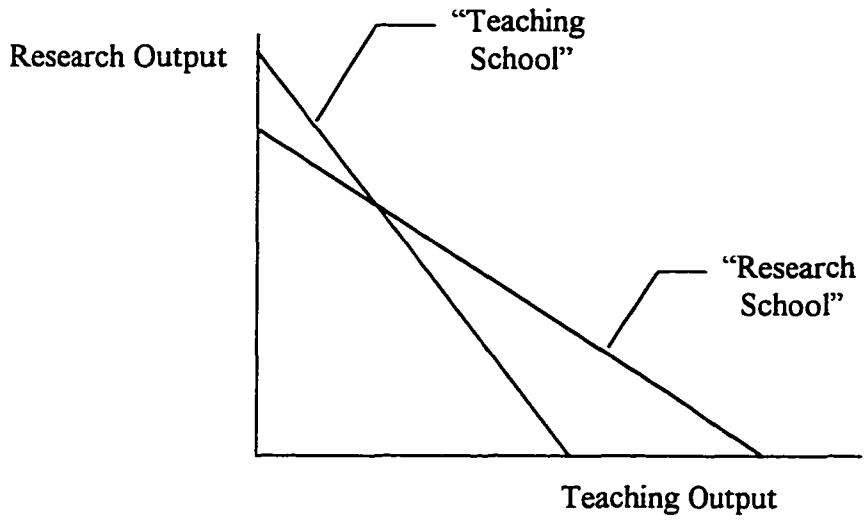


the teaching axis has a subtle interpretation and requires elaboration. That is, the teaching axis is measured from the *perspective of the students and administration*, not the faculty. To clarify, most schools have a standard number of credit hours that all full-time faculty are required to teach. But from the perspective of the students and administration, the output axis in Figure 1 measures a *quality* dimension as well (Kipps 1975). Therefore, a movement to the right on the horizontal axis can represent either more hours taught *or* more learning in the classroom with no increase in hours taught.

Broder and Taylor provide an explanation for the movements along the frontier. Although these movements are voluntary from the perspective of the faculty member, these movements respond to a set of administrative priorities. The priorities are the administrative valuation of research outputs versus teaching outputs, and these priorities establish the "relative price ratio" between outputs. This price ratio can be implicit, as

Figure 2.

Price Ratios in Faculty Performance Measurement

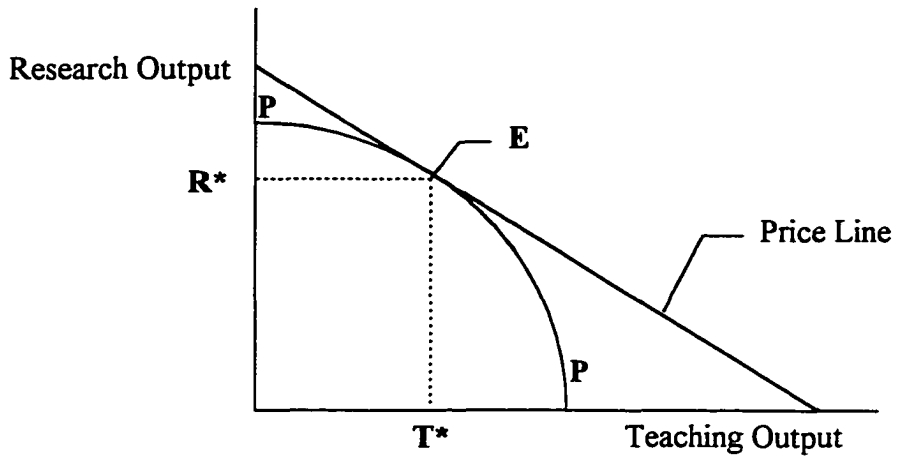


when an administration is reputed to value either research or teaching heavily, or the price ratio can be explicit, as when an administration offers "release" time from teaching with the expectation that the faculty member produce some identifiable research output. In all cases, these relative "prices" shape the economic incentives of the faculty. Again, according to Broder and Taylor, accurate measurement of faculty outputs are essential to the identification of the endpoints and the slope of the price line. This is shown above in Figure 2.

As shown in Figure 2, the flatter price line, labeled "Research School," reflects a large research priority of the administration. This price line would be characteristic of any of the major research universities throughout the country. The steeper price line, labeled "Teaching School," is characteristic of the many colleges and universities that have a primary mission of teaching.

Figure 3.

Faculty-Administration Resource Optimization



Merging Figures 1 and 2 shows the equilibrium that emerges from the interplay of faculty production possibilities and administrative output priorities. The preferred point from both a faculty and administrative perspective occurs at the tangency of the production possibilities frontier and the price line. At this point, administrative priorities and faculty outputs are optimized and the marginal rate of transformation (MRT) is equal to the slope of the price line. This tangency is shown as Point E in Figure 3. Equilibrium research outputs are at a level of R* and teaching outputs are at a level of T*.

The last step in this model is an explanation of the reward structure. Rewards in this industry take the form of merit pay, promotions, or other forms of recognition (e.g., teaching awards). The reward structure is identified by faculty attainment, or lack of attainment, of the desired administrative outcomes. If measured teaching and research outcomes are greater than the amount expected by the administration, rewards would be the appropriate economic response. This is identified by an instructor who performed

outside of the production boundary, like point A in Figure 1. Conversely, if a faculty member were performing below expectations, withholding rewards would be appropriate. An example of this is a faculty member who had an output combination inside the production frontier such as point U in Figure 1. Clearly, the system of allocation and rewards developed here is predicated on the accurate measurement of the faculty outputs of research and teaching effectiveness.

1.2. Models of Faculty Behavior

Models of faculty behavior in institutions of higher education have been developed around the labor supply paradigm of Yunker and Marlin (1984). These authors model faculty behavior and incentives in a utility maximization framework. This maximization framework envisions a two-step process. First, a typical faculty member has a conventional utility function that is constrained by the amount of money income available. The amount of income is, in turn, directly related to the workplace rewards of the faculty member. Thus, the model is characterized by an examination of the faculty member's incentives and behaviors in the workplace.

On the job, the faculty member makes cost-minimizing choices to effect the research and teaching outputs described in the previous section. Namely, the faculty member produces the combination of outputs that is consistent with the equilibrium developed earlier. This choice is consonant with the microeconomic principle of setting the marginal rate of transformation equal to the price ratio and this optimization process will be common to all faculty members.

What is unique among faculty members are their individual steps to achieve cost minimization. That is, all reward-motivated faculty members will aspire to get on, or exceed, the production possibilities curve because this curve reflects the locus of acceptable output levels. To get to higher values in output space, or to move "northeast" in Figure 1, they are required to produce more research, more teaching, or both. In the case of research, this means developing better writing or computing skills, finding better research topics, or other improvements. For teaching, this result can be accomplished in one of two ways. The first way is to improve classroom performance. An instructor can accomplish this by observing "master" teachers, preparing more extensively outside of class, attempting to use new teaching technologies, or many other possible improvements. Second, the instructor might, intentionally or unwittingly, debase the content in an attempt to curry favor with the students. This would generally lead to higher teaching evaluations, and subsequently move the instructor to a higher *apparent* level of teaching output. This debasement can occur by an overall easing of the grading standards, commonly called *grade inflation,*[19] or by changing the material in a way that the more easily-learned concepts are the majority of the topics taught. Grade inflation has been the topic of much research, and Michael Everett (1977) examined other forms of content debasement. Everett has made a convincing argument that lower-level cognitive (LLC) skills are easier to teach and learn than higher-level cognitive (HLC) skills. Examples of LLC skills are the learning of basic facts, definitions, and institutional history, while HLC skills are problem solving, discussion, and analysis. If LLC are easier to teach and learn, Everett

[19] *Grade inflation* is defined as an increase in the average grade given in a class over time when student learning and student quality are held constant.

argues that the content will be debased as faculty seek out ways to increase their measured performance.

At the core of this labor supply model is the notion that accurate measurement of teaching performance is essential to the attainment of optimal economic outcomes, because grade inflation and content debasement are less desirable outcomes than teacher improvement strategies.

1.3. Models of Educational Outcomes Using Production Functions

The dominant tool that economists have used to model learning is the *production function*. This tool is summarized succinctly by one of its main proponents, Eric Hanushek (1986, p. 1148):

> Studies of educational production functions (also referred to as "input-output" analyses or "cost-quality" studies) examine the relationship among the different inputs into and outcomes of the educational process. These studies are systematic, quantitative investigations relying of econometric, as opposed to experimental, methods to separate the various factors influencing students' performance.

Hanushek goes on to trace the history of the use of productions functions in educational research (p. 1149):

> The history of educational production function analysis is typically traced to *Equality of Educational Opportunity*, or, more commonly, the "Coleman Report." The Coleman Report was mandated by the Civil Rights Act of 1964 and was conceived as a study of the distribution of educational resources within the United States by race or ethnic background. The study, however, went far beyond simply producing an inventory of school resources. It created a massive statistical base containing survey information for more than one half million students found in some 3,000 separate schools that was employed to ascertain which of the various inputs into the educational process were most important in determining the achievement of students.

Hanushek notes that although there is disagreement about some of the technical issues involved, the use of production functions in educational research is sufficiently widespread to validate its usefulness.

If one accepts the model of a production function to be a reasonable first approximation to what happens in the classroom, several more technical issues need to be addressed. Three key issues are: the distinction between fixed and variable inputs; the selection of the appropriate input and output variables; and, the identification of production function "shift" variables. These issues are addressed below.

1.3.1. Fixed and Variable Educational Inputs

Basic production theory identifies fixed resources as those inputs that are not changeable in the short run. Typically, these are considered to be "capital" inputs. In an educational production function, the capital inputs are those student, faculty, and institutional inputs that are, essentially, fixed in the short run. Examples of student capital are native intelligence, grade point average, age, and gender. Examples of faculty capital are native intelligence, degrees attained, years of experience, and gender. Finally, institutional capital inputs are the length of the class, the time-of-day the class is offered, class size, and other fixed characteristics of the educational setting. In an educational production function model, therefore, the fixed inputs are those features of the classroom that cannot be changed during the semester the class is taken.

Variable educational inputs, on the other hand, are those inputs that *can* be changed during the semester. Stated differently, these are the variables that are within the control of the faculty and students. For the faculty, examples of variable inputs are class

preparation effort, grading behavior, and punctuality. As for the students, variable inputs include study effort, note-taking effort, attentiveness in class, homework assignments returned, and so forth.

The distinction between fixed and variable inputs to the educational production function is essential for prudent administrative decisions. Clearly, there are classroom behaviors and conditions that the faculty *can* control, along with many they *cannot* control. Any administrative reward scheme should be directed at the faculty's use of their *variable* inputs.

1.3.2. Selection of Input, Output, and Shift Variables

Having identified the key inputs and outputs of the production function, the next step is to identify those variables that are included in the model. The inclusion or exclusion of variables is based on a priori expectations, namely: Does the input variable seem likely to explain the output?

For the cognitive learning model, the determination of the appropriate specification has centered around student ability, student effort and teacher ability. These are the variables that consistently explain student learning in repeated testing.

As for the affective learning model, the principal determinants appear to be the students' feeling of being successful in the class. This success is usually reflected in the grade the students earn in the class. Because most surveys of affective outcomes are measured before the semester is completed, expected grades serve as a proxy for the students' earned course grade.

Various educational forces cannot be correctly categorized as input variables, although they influence the learning output. These variables are rightfully designated as *shift variables* or *control variables*, insofar as they shift the production function (Borg et al. 1989). An example of a shift variable is the institutional variable "night class." This binary variable separates night classes from day classes to identify any particular influence night classes might have on learning outcomes. Conceptually, nothing inherent in a night class should affect cognitive learning. Nevertheless, the night class variable might identify the influence of non-traditional students, the influence of student fatigue or other influences that may shift the production function.

The empirical studies that are cited below identify the many variables that have been used in an attempt to explain student learning outcomes.

2. Measurement of Teaching Effectiveness Using SETs

The literature reviewed in this section represents studies in the field of economic education that have measured the student evaluation process. It is well known that the use of this evaluation tool has become widespread over the past twenty-five years. Because the use of SETs is one of the several means of measuring faculty performance, careful scrutiny of the SET process has been undertaken by researchers. For the purpose of this paper, the issue of primary interest is the relationship between the unadjusted SET score and non-teacher influences upon that score. Stated differently, the question arises: What are the adjustments, if any, that must be made to the raw SET score to ensure that inter-instructor comparisons will not be biased by influences outside the instructor's control? The articles identified below have attempted to control for influences on the SET

score that might bias an inter-instructor comparison. Before the individual studies are reviewed, it is helpful to identify the *affective learning domain* and its connection with SET scores.

2.1. The Affective Learning Domain

Outputs from any educational process can roughly be divided into the affective and cognitive learning domains. As mentioned earlier, cognitive outputs are content-area knowledge outputs, whereas affective outputs are attitude-related outputs. The definition and importance of the affective domain is summarized by Phillip Saunders (1990, pp. 63-64):

> If we focus on student learning of the type of material typically presented in college level principles of economics courses, however, it is useful to formulate our objectives in terms of *both* what educators have called the "cognitive domain" and the "affective domain."
>
> The affective domain deals with feelings and emotions such as interest, attitude, and appreciation. Examples of affective objectives are: listens attentively; completes assigned homework; participates in class discussions; shows interest in economics; appreciates the importance of economics in everyday life.
>
> Much of traditional learning theory research has focused on cognitive behaviors, but, as we will see, the importance of motivation in human learning implies that we must not ignore the affective domain if we want our students to acquire, retain, and use cognitive skills.

Other authors, too, have argued that this distinction is crucial. Siegfried and Fels (1979) regard these two outputs as the most important learning areas of higher education.[20] Since affective outputs are valued by administration, faculty, and students, it

[20] Siegfried and Fels (1979), in their often-cited review of economics education research, also identify a third measure of faculty output: the number of majors attracted. Unfortunately, this measure is much more difficult to identify than the other two measures.

has become acceptable to identify and reward the production of these outputs. Both Lima (1981) and Machina (1987) have identified recommended the SET score as a suitable measure of affective outputs, and SET scores have been chosen as the measure of affective output in this study.

2.2. The Student Evaluation Process

All modern educators are familiar with the pervasive scope of student evaluations of teachers. From a social perspective, this evaluation process was partially a result of attempts by students to hold the professoriate more accountable for their classroom performance (Centra 1993). From an economic perspective, there are several reasons why SETs have flourished. First, SETs provide information about the professor that is generally available only to the students. Second, SETs provide an incentive for faculty to perform their duties. And third, SETs are a relatively cheap and easy-to-collect source of data to measure faculty performance.

Unfortunately, SETs also may produce a *moral hazard* problem (Dilts and Fatemi 1982). In general, a moral hazard occurs when two situations are present. First, economic agents obtain protection against an unpredictable and costly occurrence. Second, the existence of this protection causes the economic agents to behave, either intentionally or unwittingly, in a way that *increases* the probability that the costly event will occur. The purchase and use of insurance is a commonly cited example of a situation that induces a moral hazard problem. Applied to higher education, this problem occurs when instructors, in the interest of favorable class ratings, change the grading scale, the course content, or topic coverage in the interest of generating more favorable SETs. The

moral hazard occurs when the protection to the administration and students from derelict instructors has produced an outcome that increases the probability of that dereliction. Cognitive educational outputs may be compromised as instructors attempt, either intentionally or unwittingly, to increase SET scores.

Thus, three themes emerge from the student evaluation process. First, SETs have become commonplace among college and universities throughout the country because of the legitimate and useful information they provide. Second, the extraneous influences on SETs should be controlled to facilitate their use for inter-instructor comparisons. And third, the use of SETs as a measurement tool must be weighed against any unwanted behavioral changes that result from their use.

2.3. Variables Which May Influence SET Scores

As mentioned in the production function discussion above, the variables that predict student evaluation score have been estimated many times in the recent past. As a rule, these estimations are undertaken to identify the influences of student, faculty, and institutions on SET scores. Raw SET scores can then be adjusted to distinguish those factors that are within an instructor's control from those factors that are not. This allows for more accurate inter-instructor comparisons of their affective outputs in the classroom. A useful way to organize this literature review is chronologically.

One of the first articles to appear in the economics journals on the student evaluation process was written by Nichols and Soper (1972). This article was concerned with the connection between a student's expected grade and their evaluation of the instructor. It was hypothesized that, since expected grades are good predictors of final

grades, a student's expected grade would be highly correlated with their SET. Using an OLS estimation technique on class averages for 339 sections of social sciences classes, they found that expected grades and time-of-day variables were the only statistically significant regressors. They also found that class level and class size were insignificant. Since expected grades can be manipulated by the instructor, the authors' conclusion was the expectation that the SET process will produce grade inflation.

Kelley (1972) estimated an educational production function with the intention of finding the determinants of student evaluation levels. His OLS results found that only the expected grade was regularly significant among the student and institutional variables in his model. He also found that gender of the student, class attendance, and course requirement were statistically insignificant.

In 1973, Mirus published conclusions that were very similar to Kelley's. Namely, Mirus found that expected grades were by far the most important influences in the estimation of an instructor's SET scores. The only other significant variable from his OLS model was class size, which had a statistically significant coefficient but small effect on SET scores. The time-of-day the class was offered, and whether the course was required or elective were not significant.

In 1975, Rose attempted to control SET scores for extraneous influences. His study controlled for institutional factors and specific teacher attributes that may influence SET scores. His OLS results indicated that only 300-level classes awarded SET scores that are different from the rest of the sample. Thus, he introduced a procedure that would

correct for class level before inter-instructor comparisons are made. The percentage of majors in the class and the course requirement status did not influence student evaluations.

Dilts and Fatemi (1982) were the first authors to broach the moral hazard possibilities that the use of SETs may cause. To test this hypothesis, they estimated an OLS model with the explanatory variables of expected grade, class size, gender, and course requirement. Their empirical results supported the model they developed insofar as expected grade was positively related to SET, but none of the other variables were statistically significant.

Seiver (1983) used a Two-Stage Least Square (2SLS) estimation technique and he hypothesized that SET scores are predicted by expected grades plus several control variables. This specification suggests that, in addition to possible grade inflation, students who do well in a class are actually learning more. This logic, which is quite different from the moral hazard position, suggests that faculty are monitored by the SET process, and they produce more educational outputs *because* of the monitoring. Seiver's results suggest that SETs and expected grades are related in the usual way and that much of the connection between expected grades and SETs is because successful students learn more and therefore, reward their teachers.

Manahan (1983) published one of the few studies in the literature that has measured both affective and cognitive outputs. He used an attitude survey as his measure of affective output, and the TUCE[21] as the measure of cognitive output. Manahan used both OLS and 2SLS to estimate production functions. His attitude regressions reported that

[21] TUCE, or the Test of Understanding in College Economics, is a nationally-normed principles of economics aptitude test.

male gender and expected grades were statistically significant. His cognitive regressions suggested that age and ACT[22] scores were significant, but attitude was not. Conversely, his attitude regressions suggested that expected grades and male gender are significant. His conclusions were that cognitive outcomes are more likely to change at the completion of the class than affective outcomes.

Nelson and Lynch (1984) also evaluated the relationship between expected grades and student evaluations, and they included a discussion of the influence of SETs on grading patterns as the instructor's real income changed over time. Like Manahan, Nelson and Lynch anticipated that students who do well in a class reward their instructor, but they learn more, too. The authors test this hypothesis using both OLS and 2SLS models. Their OLS results made the oft-repeated finding that SET and expected grades are statistically significantly related. Other significant variables were that liberal arts students gave higher faculty ratings, and Saturday class students gave lower ratings. Their 2SLS results also found a positive relationship between expected grades and SETs, but it was not statistically significant. The differences between the OLS and 2SLS model are small, but the authors preferred the 2SLS specification because of its more robust results.

DeCanio (1986), in a move toward greater statistical precision, used both OLS and a multinomial logit approach to estimate a relationship between SETs and several explanatory variables. His rationale for using multinomial logit was a theme that has become more popular as this literature has matured. That is, the dependent variable is *discrete*, and OLS models assume that the dependent variable is continuous. Elaboration

[22] ACT is the common acronym for the American College Testing Program's aptitude exam.

on this distinction will come later, but for DeCanio's purposes, he used a discrete choice statistical model to estimate SET scores. His results find (p. 171), "the pattern of signs and the statistical significance of the coefficients are similar in the OLS and multinomial logit specifications..." Specifically, the significant variables are the teacher's amount of class preparation and the teacher's organization. Surprisingly, expected grade is not statistically significant, although it did have a positive coefficient. Also, class size, grade point average, and major are not statistically significant. Perhaps the reasons for the lack of statistical significance in his study is the nature of the survey instrument. In DeCanio's case, SET was estimated using many teacher attributes, for example, preparation and accessibility, instead of student and institutional attributes. Thus, the test develops a model for those attributes that students like in an instructor, but is less discriminating about whether a particular instructor is effective or not.

Gramlich and Greenlee (1993) were two of the first economists to explore the relationship among SETs and student achievement. They acknowledge that both of these outcomes are important to administrators, faculty, and students. To examine this relationship, they use both OLS and ordered probit regressions to test whether a teacher's average SET score can be used to predict the final grades students earn. Several control variables were included in these regressions. Gramlich and Greenlee's results indicate that higher SET scores predict higher final grades. The significant control variables are standardized aptitude tests, student grade point average, male gender, and student age. Even though the results were statistically significant, the authors cautioned that the results

are weak and there appears to be no strong relationship between SET scores and student learning.

A recent study that examines SET scores in the context of economics education was written by Stratton, Myers, and King (1994). The objective of this study was to examine the possible moral hazard problem, wherein instructors change their classroom behavior when student evaluations are used. The authors take data from both pre-SET and post-SET regimes, and use an OLS technique to estimate the relationship between students' final grades and the class average evaluation of the teacher. Statistically significant, positive relationships are found in the control variables of age, grade point average, repeated class, teacher's experience, and ACT. Significant negative variables were reported for female, night classes, and non-white students. The authors conclude that SETs do create an upward tilt in the grading pattern, but they caution that learning and student quality characteristics may have simultaneously improved, too.

The conclusions to be drawn from these many articles about the student evaluation process leads one to believe that the upward pressure on grades is both a theoretical possibility and an empirical reality. The difficulty with blaming SETs for grade inflation, however, lies in the simultaneous motivational and monitoring value of SETs. Furthermore, student strategies, like dropping classes more readily, may account for some of the reasons why student grades have been improving over time. Also, it is confirmed that student grade expectation, whether based on easier grading standards, content debasement or improved learning, is the most consistent and significant of all the explanatory variables.

The next section of this literature review addresses the measurement of student learning outcomes.

3. Measurement of Teaching Effectiveness Using Achievement Tests

Educators are interested in evaluating teaching effectiveness by measuring how much students have learned in addition to measuring teaching effectiveness from student opinions. Such an evaluation, however, is easy to imagine but harder to measure. The difficulty in measuring student learning is due to reasons such as: (1) the teachers may teach, but the students may not be motivated to learn; (2) the test instrument may be faulty; (3) the learning may not be apparent until a later time; (4) the teachers may "teach to the test," and only prepare the students in a narrow range of material; (5) the academic freedom of the instructor may be compromised when a common core of learning is required; (6) pre-tests are often required to isolate the learning in a particular class; and (7) student motivation to perform can be inconsistent. There are other reasons that could be added to this list. Nevertheless, student cognitive learning is an essential part of higher education, and much effort has been expended in its measurement. What follows is a general discussion of cognitive learning and the review of several articles in the field of economic education that have attempted to measure student learning.

3.1. The Cognitive Learning Domain

The cognitive learning domain has been the most-often measured outcome in educational processes and, in some contexts, has been viewed as the primary outcome of schooling. A succinct definition of this learning domain can be found in Saunders (p. 63):

> The cognitive domain deals with intellectual outcomes such as knowledge, understanding, and thinking skills. Examples of cognitive objectives are: defines basic terms; interprets charts and graphs; recognizes logical fallacies in reasoning; predicts the outcome of an action involving economic principles. Benjamin S. Bloom and others published the *Taxonomy of Educational Objectives: Cognitive Domain* in 1956. This book describes six cognitive categories in detail and presents illustrative objectives and test items for each category. The categories in ascending order are: (1) knowledge, (2) comprehension, (3) application, (4) analysis, (5) synthesis, and (6) evaluation.

Clearly, the cognitive domain embodies much of what educators call "learning." Since this learning is highly valued by administrators, employers, graduate schools, and others, it is necessary to measure cognitive learning outcomes in some reliable way. In the economics literature, the measurement of cognitive learning has proceeded much like the measurement of affective learning. Namely, economists have used a production function approach to model the learning process. Examples of this research are reviewed below.

3.2. Variables Which May Influence Achievement Exams

As with affective outputs, the production function is the common theme among the cognitive learning or "outcomes" literature. Since the early work by Hanushek mentioned earlier, economists have attempted to model student achievement as a production process. The relevant output is a student's score on a standardized exam, and the inputs are student ability and effort, instructor ability and effort, and the institutional characteristics of the course. Because the development of this literature has been cumulative, it is discussed below in chronological order.

Weidenaar and Dodson (1972) published an early study in the economics literature that addressed the subject of student cognitive outcomes. The authors used the popular

TUCE test as the dependent variable, and modeled the TUCE score as a function of student and institutional inputs. It should be noted that this specification implicitly assumed that the instructor *does not* matter in the learning process. Their OLS results indicate that significant positive influences on student learning outcomes are: age, previous college-level economics classes, ACT, and the pre-TUCE test score of the student. A statistically significant negative effect was estimated for students having a business major. When instructors variables were added to their model, it was demonstrated that faculty capital *does* matter. This result was shown by positive and significant coefficients on instructor's education, years of experience, and instructor's score on the TUCE exam.

Tuckman (1975) used both cognitive and affective measures of output to examine whether graduate student instructors performed differently from full faculty members. Tuckman used the post-TUCE score and final grades as the dependent variables in separate OLS regressions. His results indicated that significant positive explanatory variables are grade point average, instructor's experience, pre-TUCE, whereas female gender had a significant negative effect on cognitive learning. Tuckman concluded that graduate instructors have a greater positive effect on student attitudes whereas, instructors with faculty rank have a greater positive effect on cognitive outcomes.

Marlin and Niss (1980) used a production function model to test whether or not faculty inputs are important to several measures of student learning. They used a canonical correlation approach to examine whether certain faculty and student characteristics are systematically related to student cognitive outputs. The highest

correlation is between "student attributes" and outputs. In general terms, this means that the attributes of grade point average, ACT, age, and pre-TUCE are strong predictors of the outputs of TUCE-improvement, course grade, and scores on complex economic questions. The use of canonical correlation reduces the amount of precision (or *apparent* precision) between inputs and outputs in a production function. Nevertheless, Marlin and Niss defend this general approach as being more reflective of the complicated relationships characteristic of the learning process. They conclude that (p. 24), "while most of the learning is associated with student inputs, some part is due to the teacher."

Watts and Lynch (1989) examined the instructor's native language, the length of the course, the learning rate of freshman, and the effect of alternate textbooks in their study. They specified the customary production function, and used both OLS and ordered probit estimation techniques on the TUCE exam and final course grades, respectively. The relevant findings, for the sake of this research, are that student aptitude, male gender, and upper class rank are positively and statistically significant predictors of a student's achievement. Non-native English language speaking instructors and freshmen are negatively related explanatory variables. Watts and Lynch also report that faculty members are not systematically better instructors than teaching assistants.

Brasfield, Harrison, and McCoy (1993) employ an ordered probit approach to determine whether students who completed a course in high school economics perform better in college-level economics classes. Student final grades were used as the measure of cognitive outputs. The authors argue that this output measure may be preferred to an achievement exam because final grades more accurately approximate the attainment of

course objectives. Their results suggest that grade point average, ACT score, high school economics, other college-level courses in economics, and class average grades are significant explanatory variables. Gender, hours of study per week, and extracurricular activities are insignificant in their regressions. The authors note that the results on their key variable, high school economics, runs counter to the usual result in the literature which finds that high school economics is not a significant predictor of college-level success.

Two of the most recent contributions to this literature are made by Lopus and Maxwell. Their 1994 study was directed at the success in college of student who had taken high school economics, with special attention to the topic coverage in the high school class. Their dependent variable was the TUCE exam score, and they used an OLS estimation technique. The results of their regressions indicated that student aptitude and grade point average are the most compelling of the explanatory variables. Male gender is also positive and statistically significant. Among the variables that are not statistically significant predictors of TUCE outcomes are the specific instructor, hours attempted by the student, high school economics, class size, non-white students, and hours studied per week. The conclusion of the authors was that high school and college economics classes contain significantly different content.

Lopus and Maxwell (1995) also used an education production function to study the effects of different sequences of micro and macroeconomics. Again, the authors use the TUCE exam as their dependent variable, and an OLS estimation technique. Like the previously cited article, grade point average and calculus background (a measure of

aptitude), male gender, "student interest," and white students are uniformly positive and significant predictors of TUCE scores. Surprisingly, the student effort variables of attempted hours, study hours per week, and work hours per week are all statistically insignificant. The conclusion of this study was that student capital is the most important of academic success, although it should be noted that this study did not control for the instructor, but instead controlled for the type of school (research, liberal arts or two-year college).

4. Other Suggested Faculty Ranking Schemes

A final step in the literature review is to examine research that has developed non-traditional faculty ranking schemes. Six studies in the economics literature are reported here. These studies have all constructed *adjusted* faculty ranking schemes. Because the sophistication of this research has seemed to improve over time, it is reviewed in chronological order.

Rose (1975), as discussed earlier, devised a "first-generation" faculty ranking scheme to acknowledge the frequently stated observation that SETs reflect class conditions over which the instructor has no control. His scheme is based upon controlling SET scores for the *extrinsic* influences of class size, class level, percentage of majors in the class, and whether the class was required or not. Using class averages of the above-mentioned variables and an OLS estimation technique, he found that only 300-level classes exert a statistically significant influence on SET scores. Rose's method is used to adjust the raw instructor rankings for the extrinsic influence of class level. His adjusted rankings

of professors resulted in a small, one or two position, change in the ranks of 13 of the 36 instructors in the sample.

Dilts (1980) addressed his study to the negative effects of grade inflation that were becoming evident in the late 1970's. To counteract the temptation to inflate grades to increase SET scores, Dilts proposed an instructor ranking scheme that corrects for grade inflation. Dilts used OLS to estimate a SET equation, and found that expected grades and required classes are the only predictive independent variables. Although Dilts did not explicitly rank professors, he did argue that raw SETs should be adjusted downward if a professor's grading pattern is higher than the departmental average. Additionally, raw SETs should be adjusted upward in a situation where a professor taught a higher than departmental-average number of students taking the class as a requirement.

Zangenehzadeh (1988) also directed his study to the unfavorable effects of grade inflation in schools of business. He contended that the value of grades should not be sullied by the manipulation of instructors because the grades of the student serve an important screening service for employers, graduate schools, and the students themselves. To identify the relationship between grades and SET scores, Zangenehzadeh uses both OLS and Three-Stage Least Squares regression techniques. His results suggest that SET scores are positively, and statistically significantly, related to a students' expected grade and the general "quality" of the course. His proposed adjustment is straightforward: compare the departmental average grade with an individual instructor's average grade, and adjust the SETs accordingly. This proposed scheme is applied to 39 business instructors.

Although most of the rankings were changed somewhat, no statistical tests are applied to identify whether the ordering could have been changed as much by random chance.

Schmanske (1988) offered a faculty ranking scheme based on a combination of SET scores and cognitive outputs. He posed the question: is the relative class performance of a student in a *second* economics class predicted by the instructor who taught the *first* class? If the relative class performance in the second class systematically improved from the first class, some of this improvement should be attributable to the first instructor. From these assumptions, Schmanske uses OLS to estimate a learning equation, and finds GPA and relative performance in the first class to be statistically significant predictors of second class success. Schmanske also includes individual instructor dummy variables. Although none of the individual teacher coefficients are statistically significant, these coefficients are used to rank the instructors. The result is a positive, but statistically insignificant relationship between the SET ranking and the cognitive ranking. Schmanske concludes that this study provides weak support for the use of SETs as measures of teaching effectiveness.

Two recent studies, by Mason, Steagall, and Fabritius (1995), and by Zietz and Cochran (1996), make further refinements on an instructor ranking scheme. Mason, et al., apply an ordered probit regression technique to SET scores in an attempt to establish whether they are influenced by teaching, course, or student characteristics. Their results identify several characteristics that statistically explain SET scores. Most notable among these are expected grades, male gender, and student effort. After adjusting the raw SET scores for those outside influences, they re-rank the instructors. Although there are

several instances of large changes in rankings, the new scheme has a Spearman rank correlation coefficient of .92 with the raw rank ordering. In spite of the high correlation between the old and new, they conclude (p. 414), "the validity of raw rankings of faculty members for the purposes of promotion, tenure, and raises should be seriously questioned."

The final article to be reviewed, by Zietz and Cochran, has created much of the motivation for the research contained herein. Their model, like the previous model developed by Schmanske, collected student data from both cognitive and affective learning outputs. The cognitive output was the TUCE exam, and the affective output was the instructor's SET score. The data for their study were taken from the norming session of the TUCE exam in 1989-90, and they use a random-effects regression model to estimate both a SET and learning equation. By adjusting the raw SET and learning equations for influences outside the control of the instructor, an adjusted ranking scheme is created. Their results suggest that the differences between the raw and the adjusted instructor ranks are significant, which elicits the conclusion that the traditional technique of comparing an instructor's raw SET score to the departmental average is deficient. The innovation of this research is to merge the SET score with an achievement score to form a composite measure of teaching effectiveness. Once this measure has been constructed, a weighting scheme can be developed by the administration to reward the outcome they value most highly. In the limit, an administration could value either outcome exclusively. More likely, however, an administration values both outcomes (as suggested by Saunders above), and the weighting scheme is related to the mission of the school.

5. Chapter Summary

This chapter has focused on several key themes in the economic education literature. First considered were those theoretical models that addressed the resource allocation problem faced by college administrations. Next, faculty reward and motivation models were considered. It was concluded that the measurement of faculty output is essential to any well-functioning system of allocation and reward.

The second key theme reviewed in this chapter addressed the empirical model that has conventionally been applied to educational research. That model, the production function, was then used to estimate the relationship between various inputs and outputs of cognitive and affective learning. The literature suggests that cognitive outputs are chiefly explained by teacher quality, student aptitude, and student effort. Affective outputs are primarily explained by a student's success in the course, and this is usually reflected in their expected grade.

Finally, this chapter considered several alternative ranking schemes that have been proposed in the economics literature. The general tendency of these schemes is to adjust output measures for those factors that are outside the direct control of the instructor. In other words, the alternative schemes attempt to rank instructors on their measured variable labor inputs over the semester.

CHAPTER 3

AN ECONOMIC MODEL OF TEACHING EFFECTIVENESS

The models that form the bulk of neoclassical economic theory rely on incentives, exchanges, prices, incomes, information and other characteristics that make up a *market*. These models also assume that economic agents are rational and purposeful in their decision-making. Economic models of teaching effectiveness are of the same variety. Namely, these economic models assume that the economic agents have some explicit or implicit "objective function" and a series of constraints. In this study it is assumed that the objective function college and university administrators maximize is educational output, which consists of research and teaching quantities.[23] The other relevant group of economic agents, the faculty, are assumed to maximize their individual utility functions. The administration is subject to a budget and other institutional constraints,[24] and the faculty is constrained by their individual incomes, their native talent, and available academic resources.

Within these basic objectives and constraints, it is assumed that both administrators and faculty members have some opportunity to exchange and to allocate resources in ways that pursue the optimization of their respective objective functions. Each of these optimization procedures will be considered in turn. Thereafter, models of teaching output

---

[23] It should be noted that the faculty of some colleges and universities have no research expectations; their output is teaching only. Although these schools are disregarded in this study, the general conclusions of this paper can be extended to those schools.

[24] An example of an institutional constraint is an outside accrediting agency or a major donor.

measurement are discussed. These models provide the theoretical underpinnings of the classroom learning process and, therefore, provide the foundation for the empirical estimates of cognitive and affective learning which follow.

1. A Model of Administrative Behavior: Resource Allocation and Reward Schemes

Assuming that administrators allocate their resources to generate an optimum combination of research and teaching outputs, the following discussion explains the optimization process and elaborates on information problems which may be encountered as this optimization proceeds.

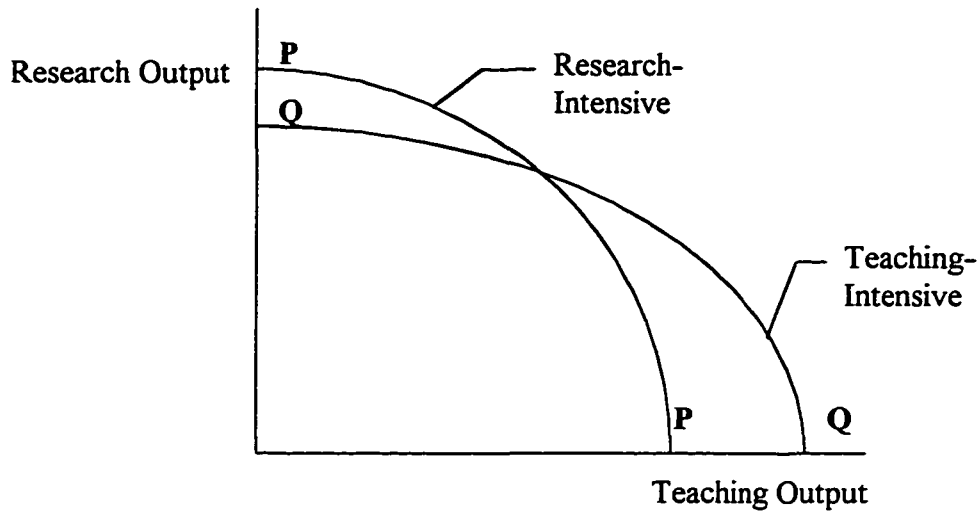
1.1. Allocation of Faculty Pursuant to Comparative Advantage

One of the most venerable ideas in economics is that resources should be used to their *comparative advantage*. Simply stated, this principle suggests that resources should be deployed where they have their lowest opportunity cost; low-cost producers should be used instead of high-cost producers. As applied to higher education, administrators evaluate the skills of their faculty and determine who are the low-cost producers of research and teaching, and deploy the faculty accordingly. But how do administrators determine who are the low-cost producers? Becker (1979, p. 1016) speaks to this problem:

> In the case of research, the existing screening methods are generally agreed upon and considered highly accurate. For example, an article in a prestigious, refereed journal is accepted as an indicator of quality output. As such, a university interested in raising research output may only need to raise the income determination weight given to research.
>
> Teaching, unlike research output, has no existing measure which is universally accepted as highly accurate....Only if a university is able to adopt student evaluations, standardized student learning measures, or some other proxy

Figure 4.

Faculty Ability Endowments



index of teaching output can the reward structure be used to cause an increase in every faculty member's desire to increase productivity in teaching.

Clearly, the principle of comparative advantage relies on the accurate measurement of faculty output and thus, there is an established need for the reliable measurement of teaching effectiveness. To understand that need in the context of this research, it is helpful to look at a series of production possibility frontiers. This examination provides the basis for an understanding of faculty comparative advantage.

In Figure 1 (page 17), a higher education production possibilities frontier was introduced. That frontier is to be understood as a representation of the *average* faculty trade-off between research and teaching outputs. Typically, however, there is variation in skills and interests among faculties, and this variation gives rise to frontiers of several other shapes. Figure 4 shows two production possibilities frontiers that are composed of distinctly different faculty members.

Figure 4 shows a faculty that is not made up of homogeneous labor inputs, but rather is composed of faculty members who differ in their native talents or interests. The "Research-Intensive" instructor, on frontier PP, is characterized as one who has the native ability and motivation to produce research outputs. The "Teaching-Intensive" instructor, on frontier QQ, has the ability to produce more teaching output but less research vis-a-vis the research-intensive instructor. Since both instructors face a trade-off between research and teaching,[25] the department's implicit price ratio determines how much of each output they will individually produce.

Assuming that a department has some variation in their faculty composition and assuming that resources can be deployed where they are most highly valued, the administrative allocation of resources will follow from the comparative advantage principle. Namely, the least-cost producers of research and teaching will be identified by their measured research and teaching outputs, and will be deployed accordingly. Accurate measurement of teaching effectiveness is essential to this allocation process.

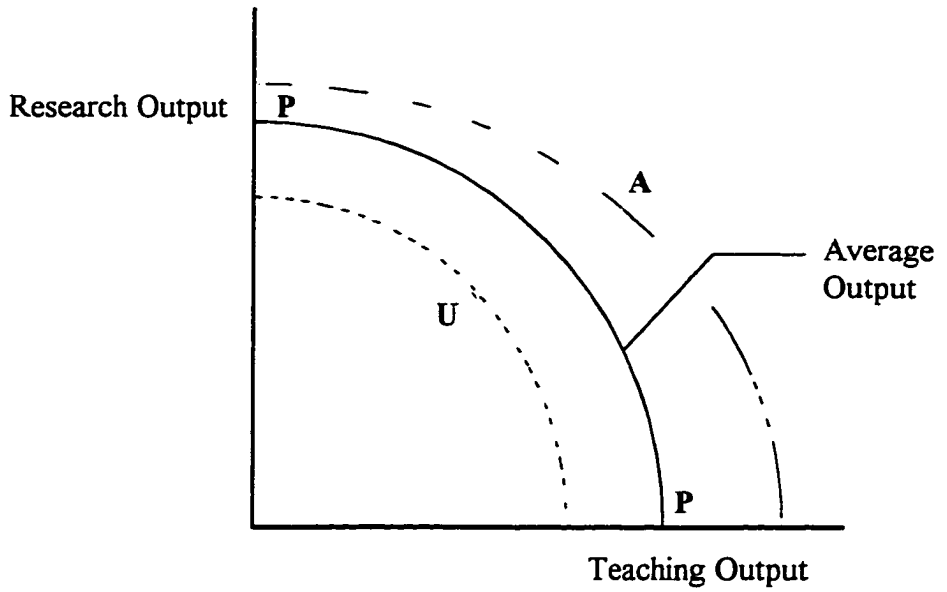
1.2. Rewarding Faculty in a Regime of Asymmetric Information

The foregoing discussion assumes that there is accurate measurement of research and teaching outputs. But what if, owing to incomplete information, the faculty has different information from the administration? The most likely combination of asymmetric information is when the faculty has more information about classroom outputs than the

[25] There is a possibility that research and teaching have *complementarities*. If they do, there would be a positively-sloped portion of the production possibilities frontier. Even though this may be possible, Yunker and Marlin (1984) argue that the negatively-sloping portion of the production possibilities frontier is the only economically relevant portion for analysis. This is because optimization occurs where slopes are equal, and the implicit price line is always downward sloping.

Figure 5.

Absolute Advantage in Faculty Skills



administration. This incomplete information situation was depicted in Figure 1 (page 17), where points above and below the production possibilities frontier were shown. The explanation given for points off the frontier was that instructors know of their output vis-a-vis the departmental average, but administrators do not have this information because they are not in the classroom. Thus, superior performers and inferior performers can coexist. The situation wherein a faculty has some members who are *absolutely* superior or inferior performers compared to the average instructor is depicted in Figure 5.

Points on the frontier labeled A in Figure 5 indicate a faculty member with more absolute abilities than the average faculty member, who is positioned on frontier PP. Similarly, the frontier labeled U indicates a faculty member who is endowed with fewer skills or less motivation than the average faculty member.[26] Clearly, the comparative

[26] A tenured contract may have the same motivational effect.

advantage allocation scheme is not of much help in finding an optimal outcome in Figure 5 because there is no price line that can efficiently allocate these resources. In this case, a reward scheme is needed. If there are ways to reward faculty members who produce on frontier A, and penalize faculty members on frontier U, then a departmental optimum can be reached.

Because the classroom environment is typically closed to outside observation, a situation of incomplete information is highly possible. This incomplete information may preclude the identification of all but average levels of instruction. In the absence of accurate measures of classroom outputs, specifically teaching effectiveness, production frontiers such as A, U and PP can coexist, and the optimum faculty reward scheme may not be realized.

2. A Model of Faculty Behavior: Utility Maximization

In this model it is assumed that faculty members, like other economic agents, have the ultimate objective of utility maximization. The following discussion explains how faculty members are modeled as suppliers of labor, and how this labor supply process results in utility maximizing outcomes.

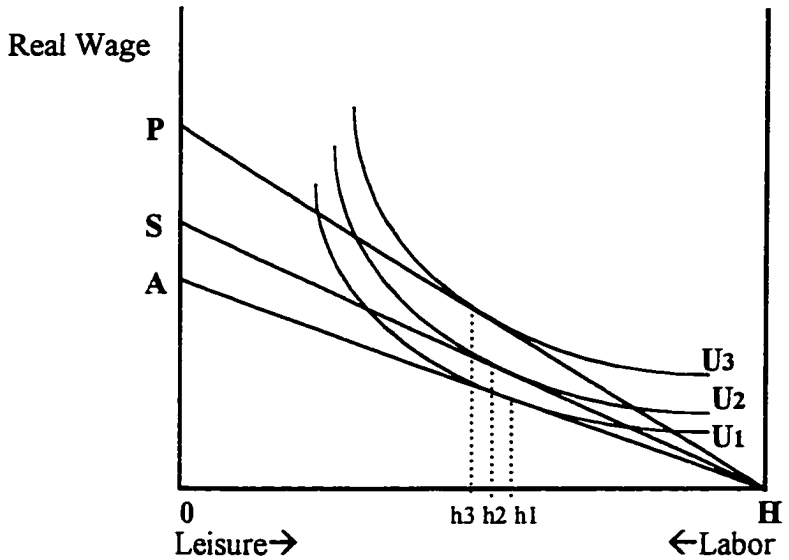
2.1. The Faculty Member as Supplier of Labor Services[27]

The job of a faculty member is to supply the labor services of teaching and research, as has been discussed several times. The quality, quantity and remuneration of faculty labor services can be modeled using the neoclassical labor supply paradigm.

[27] This section draws heavily from McConnell and Brue (1995).

Figure 6.

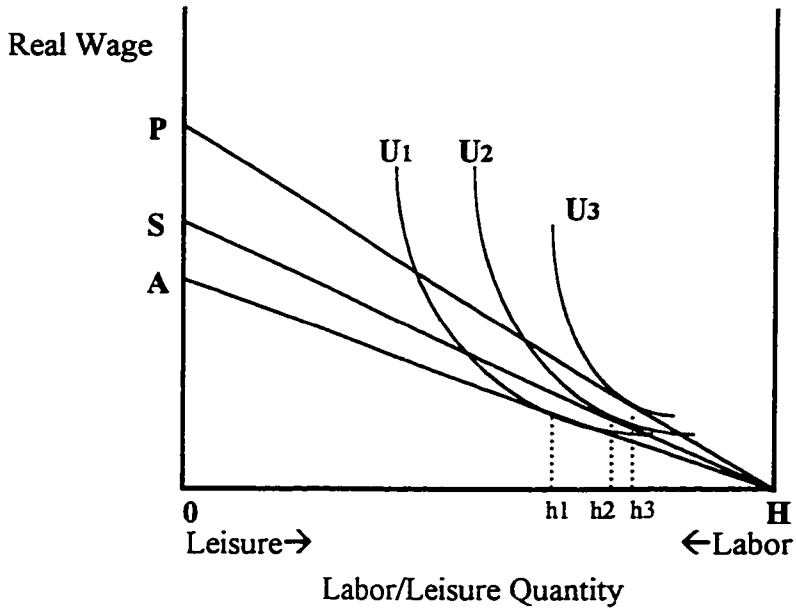
Faculty Labor Supply I



The theory of labor supply postulates a relationship between labor units supplied and real wages. This model envisions workers allocating their 24 hour day among work and leisure in response to the influences of real wages and tastes. The result of this allocative process determines the hours of labor supplied by an individual, and by extension, the labor supply curve at the market level. Unlike the typical upward-sloping supply curve for other goods and services, however, the labor supply relationship can be either positive or negative. That is, an increase in the real wage can cause *either* an increase *or* a decrease in the amount of labor services supplied, because workers face a trade-off between work and leisure. If it is assumed that higher wages induce greater amounts of labor quantities, the relationship shown in Figure 6 applies. In this case, the real wage line is rotated through three hypothetical wage levels: HA for Assistant Professor wages, HS for Associate Professor wages, and HP for Professor wages. Labor

Figure 7.

Faculty Labor Supply II



effort forthcoming is plotted on the horizontal axis below the tangencies between the wage line and the indifference curves labeled U1, U2, and U3.[28] As shown, an increase in the real wage has resulted in both an increase in utility level and an increase in labor effort forthcoming (from h1 to h3).

It is also possible to imagine a situation where an increase in the real wage causes a *decrease* in the amount of labor effort forthcoming. This situation occurs when an increase in the real wage induces faculty members to "buy" more leisure with their increased income. An example of this situation is shown below in Figure 7, where the wage line is again pivoted upward through wage levels HA, HS, and HP and the resulting quantity of labor supplied falls from h1 to h3. Technically, this outcome is an example of

[28] These indifference curves are also understood to be iso-utility curves. Utility levels increase as the curves move outward from the origin.

where the "income effect" of higher real wages has overtaken the "substitution effect" of higher priced leisure. When seen at the market level, this situation results in the "backward bend" of the labor supply curve (McConnell and Brue 1995).

Two extensions of this model are required for completeness. First, the labor quantity axis must be carefully defined. In the usual application of this model, all units of labor are considered to be homogeneous. But a slightly different interpretation is required for the faculty labor supply model owing to the heterogeneous nature of the "quantity" axis. Because faculty members produce *both* research and teaching outputs, and because both of these vary in terms of quality, the horizontal axis in Figures 6 and 7 implies *both* a quantity *and* a quality dimension. Better teaching, for example, would result in a leftward movement along the horizontal axis, even if the number of assigned teaching hours were unchanged.

Second, the trade-off between wages and leisure makes the results of any administrative reward scheme uncertain. That is, if any faculty members have high preferences for leisure as compared to income, their response to an increase in the real wage may be to supply *less* labor. This possibility complicates the reward schemes that were developed earlier where it was assumed that higher rewards would bring forth higher levels of output.

To recap, it is assumed that faculty labor supply generally operates within the neoclassical microeconomic model. The extensions and clarifications of this model are required for completeness of the theory, and also to suggest that empirical abnormalities are possible when this model is used to evaluate faculty behavior.

2.2. Methods of Faculty Resource Use

Implicit in the foregoing model is the notion that all instructors are efficiently producing the quantity of labor supplied at all levels of the real wage. But an instructor could be an inefficient producer for a number of reasons. The most important of these reasons are explained below.

First, an instructor may be inefficient because the technology used in instruction and research is outmoded. Clearly, the near-market process that exists in the research area would quickly resolve this. The non-market nature of teaching, however, may suggest that classroom technological changes are more slowly adopted.[29] Therefore, it is expected that not all classroom instructors are using the same technology, and this may limit the instructor's output.

Second, a teacher may be inefficient because he or she is a "poor teacher." To move to the efficient level of output, their teaching skills must improve. There are several ways that teacher improvement can be undertaken. Much teacher improvement seems to come from experience (Lima 1981), but it may also be true that improvement comes from better preparation, changes in the class format (as from lecture to discussion), and speech and diction improvement. Unfortunately, no teaching improvement strategy appears to be universally successful (Siegfried and Fels 1979), so it is likely that several paths to the efficient level are tried.

Third, instructors can change the course content to *appear* to improve academic

[29] Kelley (May 1972), in using computer-aided teaching techniques, has concluded that technological change can improve learning. Others, however, (Siegfried and Fels 1979; Bosshardt and Watts 1994) have reported that technology is of lesser importance.

outputs. This suggests that a change in academic output may not involve a change in delivery technique but a change in course rigor. Two strategies in this vein are content debasement and grade inflation. Content debasement makes the course material easier by tilting the content towards subjects that the students find easier to learn. An example of this in economics is the teaching of more institutional and descriptive material at the expense of graphical presentations (Everett 1981). Grade inflation is another way to give the appearance of teaching more without doing so. It has been shown (Saunders 1990) that student learning goals are shaped by the teacher's course objectives; the student's standard of learning is positively related to how much the teacher expects of the students. Therefore, if the grading scale is made easier, a student may *feel* that they have accomplished more in the class, and it will appear that the instructor has imparted more knowledge.

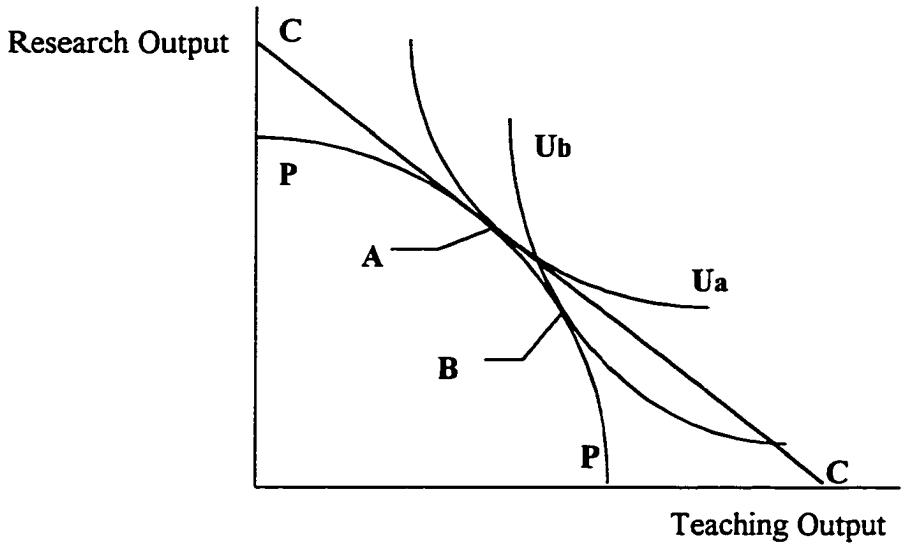
Thus, faculty members aspire to produce efficiently, and this can be achieved in several different ways. From an administrative perspective, the preferred means are increased real output by using improved technologies, improved teaching techniques, or more intensive application of current techniques. But the "increase" in teaching outputs can also be achieved by using strategies which emphasize apparent improvement rather than real improvement. It is expected that *instructors choose the cost-minimizing technique* among the several alternatives (Stratton et al. 1994).

2.3. Multiple Faculty Equilibria

Casual observation indicates that faculty produce different levels of research and teaching outputs and achieve different rewards and ranks. Accounting for these multiple

Figure 8.

Multiple Faculty Equilibria



observed outputs requires an extension in the use of utility curves as in Figure 8. In this case, the axes are labeled with the faculty outputs of research and teaching.

For the sake of exposition, one *average* production possibility frontier (PP) has been drawn in Figure 8, along with two indifference curves, labeled Ua and Ub. The representative faculty member, with tastes given by indifference curve Ua, gravitates toward point A in Figure 8. Point A is the tangency point between the indifference curve Ua, the price line CC, and the production possibilities frontier curve PP.[30] Another faculty member, because of different preferences toward research and teaching, has the more steeply-sloped indifference curve, Ub. These different preferences cause gravitation toward equilibrium point B, which contains higher teaching outputs but lower research

[30] Technically, at point A the marginal rate of transformation of PP equals the marginal rate of substitution of the indifference curve Ua, and also equals the slope of the implicit price line CC.

outputs than point A. However because B is inside the implicit price line, this faculty member has *not* achieved the minimum departmental output that is established by CC. An equilibrium at point B can only be realized if the indifference curve of professor B is *steeper* than the average professor in the area around point B *and* there are incentives embodied in the system of rewards and ranks. In a regime of merit pay and/or promotion, a faculty member in equilibrium at point B will *not* be rewarded, although he or she will have achieved a personal utility maximization. Examples of equilibrium points like B are found in faculty members who, despite considerable experience, are never promoted to the highest rank.

An example (not shown) of a faculty member that regularly performs above the departmental standard can easily be imagined. This instructor would operate on an indifference curve outside of both Ua and Ub. In a regime of merit pay and promotion, this employee would more quickly achieve the maximum rank and pay possible.

2.4. Long-Run Faculty Expansion Paths

The time allocation between research and teaching is also of considerable importance to the faculty. As Broder and Taylor (1994) explain, if teaching outputs are mis-measured, both short-run and long-run consequences are realized. Figure 9 shows the hypothetical mis-measurement of faculty teaching outputs. In this case PP is the actual production possibilities curve, whereas QQ is the improperly measured production possibilities curve where the administration has understated the maximum amount of teaching output that can be produced.

Figure 9.

Long-Run Consequences of Output Mis-Measurement

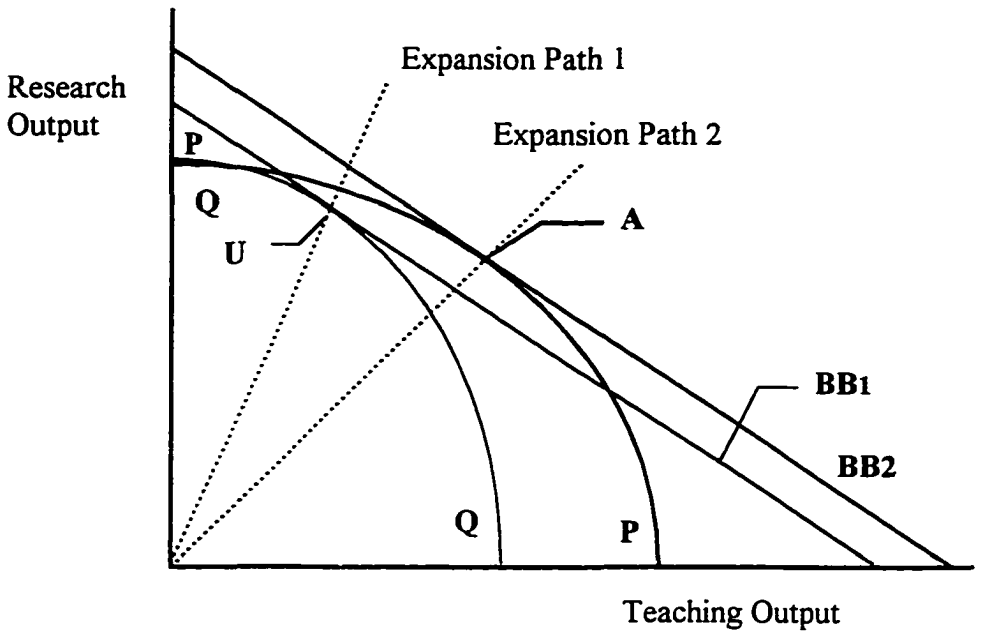


Figure 9 shows, in the short-run, that administrative understatement of potential teaching output has reduced the total output from level A to level U. That is, if the measurement of both outputs were accurate, resources would be more heavily used in teaching in the short-run, and point A would be preferred to point U. However, because the intersection of the production possibilities frontier and the horizontal axis has been understated, the equilibrium occurs on frontier QQ at a tangency to the price line BB1, instead of the parallel price line BB2. It is clear that the administration would prefer to move along frontier PP because all points on this curve are above frontier QQ. Moreover, the long-run allocation of resources is adversely affected because faculty development will follow Expansion Path 1 instead of Expansion Path 2. Faculty members will inappropriately over-allocate research outputs, and under-allocate teaching outputs.

3. Models of Teaching Output Measurement

It has been established that administrators and faculty are widely affected by measured classroom outputs. The final step in the development of the aforementioned models is to explicitly address the measurement of the teaching output axis. This involves the use of educational production functions, which develop the relationship between teaching inputs and student learning. After identifying the relevant inputs and outputs, these models are statistically estimated to measure levels of teaching effectiveness. But as discussed in Chapter 1, teaching effectiveness is a combination of both cognitive and affective learning outputs. Each learning output is discussed in the sections that follow.

3.1. Modeling Cognitive Learning Outcomes

A production function model is used to measure cognitive learning outcomes because this model provides the identification of inputs into the learning process and, in particular, the specific contribution of the instructor. Thus, it is useful to assume that each teacher's variable labor inputs are tantamount to a different level of "technology," and this suggests that each instructor operates on a unique production function. This technique helps to identify the instructor's contribution to student learning. A useful explanation of this model is reported in Siegfried and Fels (1979, p. 926):

> To evaluate innovative instructional techniques it is necessary to hold other things, especially the level of inputs, constant. William I. Davisson and Frank J. Bonello describe a useful taxonomy for organizing research on the production function for learning college economics. Their approach is superior to the *ad hoc* theorizing that characterizes most of the economics education literature. Davisson and Bonello identify three separate categories of inputs: human capital (SAT scores, grade point average, and pretest scores); utilization rates (time spent on the course by students); and technology—the efficiency with which effort is

> transformed into cognitive achievement (lecturer effectiveness, text effectiveness, *etc.)*....A shift in the production function as a consequence of some alternative technology can be detected correctly only if input rates are held constant. Otherwise, performance comparisons consist of output at different levels of inputs on potentially different production functions, and it is impossible to disaggregate the effect of changes in the level of inputs from changes in the rate at which inputs are transformed into outputs.

Stated simply, the Davisson and Bonello (1976) approach controls for the various levels of student and institutional inputs and, as such, is able to identify the contribution of each individual instructor. Because much of the literature has identified student ability and effort as key variables in the cognitive learning function, these variables must be controlled to identify the amount of learning that is rightfully attributed to a specific instructor. Other biographical (e.g., gender, age) and institutional variables (e.g., time-of-day, size of class), as well as those characteristics that are not "variable labor inputs," must also be included among the controls.

Statistical estimation of this model allows the measurement of cognitive levels of learning. Cognitive achievement, in conjunction with affective learning, are the key components in measuring the abscissa of the graphs that are found in this chapter.

3.2. Modeling Affective Learning Outputs

Affective output measurement is also estimated using production functions and the Davisson and Bonello approach. In this case the output is the SET score, which has been identified as a good yardstick of students' attitudes about economics. The literature has identified several instructor characteristics that consistently predict positive student learning attitudes. Foremost among these is the feeling of "doing well" in the class, as

represented by a student's expected grade. Therefore, it is anticipated that expected grades will explain much of the SET score. But, like the cognitive learning production functions, student biographical and institutional characteristics must also be controlled.

Like cognitive learning, statistical estimation of the affective learning model allows for the quantification of the teaching output axis in the graphs developed in this chapter. And the joint estimation of cognitive and affective learning outputs allows for the identification of teaching effectiveness as defined in this research.

4. Chapter Summary

This chapter has explored the microeconomic underpinnings of the teaching effectiveness models. The discussion centered on the two chief economic agents in higher education: the administration and the faculty. The vital importance of accurate measures of teaching effectiveness has become apparent as these models are reviewed. The chapter concludes with the identification of a student learning model which can be used to measure teaching effectiveness. The estimation of these learning equations follows.

CHAPTER 4

THE DATA AND STATISTICAL METHODOLOGY

This chapter examines the data and the statistical procedures used in this study. The first part of the chapter examines the data sources, data quality issues, and the data collected. The second part of the chapter evaluates the statistical techniques that are commonly considered when educational production functions are estimated. The chapter concludes with the selection of the statistical estimation techniques that are used in this study.

1. The Data

The statistical estimation of the production functions developed in this study rely on three major sources of data: student evaluations of the teacher, an achievement test, and administrative data about both students and the faculty. Each of these data sources are discussed more fully below.

The student evaluation of teacher (SET) data used in this study were collected during the last week of class during the Fall semester of 1996 at Middle Tennessee State University (MTSU).[31] The population of interest was students enrolled in introductory college-level economics. The sample draws data from students in twenty-four sections of principles of economics classes (both microeconomics and macroeconomics). The SET

[31] Research on "human subjects" at MTSU must be approved by the Institutional Review Board. A letter of approval for this study is found in Appendix 1.

questionnaire elicits a one-word evaluation of the instructor, and follows with several biographical questions about the student. The questionnaire was administered by a graduate student, the instructor was not present when the questionnaire was administered, and the students were offered a small reward for participation in the study.[32] Students were assured of their anonymity and told that the data were to be used "for research purposes only."[33] The survey yielded 571 responses from a total of approximately enrolled 800 students, and a large majority of the surveys were complete and usable.

The achievement test data were also collected from the same population of students, and the test was administered during the final exam week of the same semester. Separate microeconomics and macroeconomics comprehensive exams were written and used. Each exam consisted of 30 questions and followed a multiple-choice format. The tests were prepared with input from the instructors teaching the course, the instructor's manual accompanying the textbook for the course, and the author of this paper. There were several objectives in the design of the tests. These objectives included making the length of the test the same as the popular TUCE exam, coverage of topics in areas that are widely regarded as representative of the subject matter, ensuring sufficient rigor, and the inclusion of both descriptive and analytical questions in an attempt to measure both higher and lower-level cognitive skills.[34]

The final source of data was administrative records. Data collected from this source provided additional information about the academic and biographical backgrounds

---

[32] A copy of the survey questionnaire is found in Appendix 2.
[33] Hanson and Kelley (1973) have shown that students answer differently when the data are collected for merit and promotional purposes as opposed to teacher improvement purposes.
[34] A copy of the achievement examinations can be found in the Appendix.

of both students and instructors. Because some of the data were available from both sources, these data also served as a cross-check of the self-reported data.

1.1. Data Quality

In any study that is supported by statistics, it is appropriate to consider the quality of the data upon which the inferences are made. Three key areas to examine are the accuracy of the data, the reliability of the survey and test instruments, and the representativeness of the data. These issues are considered below.

1.1.1. Data Accuracy

Questions about data accuracy typically involve scrutiny of the source of the data. To isolate problem areas in data collection, a distinction must be made between the data that were "self-reported" and the data that were collected from administrative sources. If it can be assumed that the administrative data are accurate, the only source of data inaccuracies are from the SET questionnaire.

The SET questionnaire elicits a student's evaluation of their teacher and asks for several pieces of biographical information. But self-reported information is sometimes inaccurate due to misreporting. For convenience, the problem of misreporting data can be divided into "unintentional" and "intentional" misreporting. Unintentional misreporting is defined as the situation where a survey respondent unwittingly gives an incorrect answer to a question. This may be due to misreading the survey question, ambiguous questions on the survey, fatigue, carelessness, or other reasons. In this research the survey

instrument was written in an attempt to elicit short, straightforward answers. Because of the direct nature of the questions, it is assumed that they were self-explanatory and clear.

A second, and often times more problematic area of data accuracy regards the self-reported nature of the biographical data. Self-reported statistics give the respondent the opportunity to exaggerate or embellish the truth. This is a common problem in educational research. As one might expect, students sometimes report what they would like the correct answer to be, not what it is. Maxwell and Lopus report their findings about this issue (1994, p. 201):

> Our research indicates that students tend to overstate their academic accomplishments....This produces a "Lake Wobegon" effect in student self-reported data.
>
> Overstated achievement may produce biased estimates of the relationship between achievement and educational inputs, if overstatement is correlated with achievement.

Clearly, some statistical quality control is necessary. In this study, because both self-reported and administrative data were collected, a comparison of the two data sets is made. This comparison yields several interesting and useful findings. First, students in this study *do* tend to overstate their academic achievements. Assuming that the administrative grade point averages are correct, the mean student grade point is overstated by an average of .18 (on a 4.0 scale). But for non-grade data, the reporting accuracy seems to be much improved. The self-reported age variable has a correlation coefficient of .98 with the administrative value, and the self-reported gender value has a correlation coefficient of 1.0. The combination of the grade point average, age and gender reports

leads to the following conclusion: administrative data should be used when available, but the self-reported data in this study appear to be highly accurate.

1.1.2. Survey Instrument Reliability

Another area of interest in the data collection process is the "reliability" of a measured variable. Statistical reliability is defined as *the consistency of a measured variable in repeated samples* (Jacobs and Chase 1992). Clearly, if there is high consistency from sample-to-sample, statistical conclusions are less susceptible to an aberrant sample. The SET data and achievement test have both been tested for reliability, and both are judged reliable by popular statistical tests.

The reliability of the SET survey is tested by a comparison of the results of two student evaluations done on successive weeks at the end of the semester in Fall, 1996. The first SET survey was administered by the economics department, and the second survey was given for the purposes of this study. The survey instruments were nearly identical, however one noticeable difference is that the departmental SET asked several questions about the instructor whereas the "research" SET contained only one question about the instructor. The teachers are then ranked according to the results of each survey. The results of these surveys and the test statistic that compares the similarity from sample-to-sample are found in Table 1. As is clear from the table, there is a high correlation of instructor ranks from sample-to-sample. This is borne out by the high (.85) and statistically significant correlation of the SET scores according to the Spearman Rank Correlation Coefficient. It is expected that the results of these two evaluations would be highly similar since a student's opinion of their teacher's performance is unlikely to change

TABLE 1. Rank Order of Research vs. Departmental SET Score

INSTRUCTOR	RESEARCH SET RANK	DEPARTMENTAL SET RANK
A	1	1
B	2	3
C	3	4
D	4	2
E	5	5
F	6	11
G	7	6
H	8	7
I	9	9
J	10	10
K	11	8
L	12	12

NOTE: The Spearman Rank Correlation Coefficient equals 0.85 ($p < 0.001$).

greatly in a short time. The small differences in rank ordering may be attributable to a slightly different group of respondents from survey-to-survey or a change in some students' attitudes from week-to-week.

The reliability of the achievement test is more difficult to establish owing to the fact that repeated samples are not available. Nevertheless, the relationship between a student's actual class grade and the achievement test provides some guidance about the reliability of the test instrument. Using this logic, a simple correlation coefficient has been constructed to compare these two attributes of student performance. The results bear out the expected positive correlation between a student's overall performance in the class and their performance on a standardized achievement exam. The Pearson Product Moment Correlation coefficient is .54, and a z-test of this coefficient versus zero is significant at the

0.001 level. Thus, it is shown that high-performing students scored better on the achievement exam than low-performing students, even though this correlation was *not* controlled for the weight of the exam on the final grade, student aptitude, the teacher, or other variables.

The conclusion to be drawn from these tests of reliability is that *the survey instruments used in this study were consistent measures of affective and cognitive learning*. Moreover, the finding that the SET score has test-retest reliability agrees with the majority opinion in the literature "that students can rate classroom instruction with a reasonable degree of reliability" (Costin, Greenough and Menges 1971, p. 513). Similarly, the reliability of the achievement test follows from the logical connection between a single exam and a student's final class grade: in general, those students who perform well on a comprehensive achievement exam are those who receive a high grade in the class.

1.2. Data Representativeness

A final issue in the area of data quality concerns the representativeness of the data. In short, this question asks, "Do the data represent the population from which they come?" For the purposes of this research, two considerations are made in addressing this question. First, is the sample drawn in an unbiased way? And second, do student withdrawals affect the inferences that are drawn from this sample? Both of these issues are addressed in turn.

1.2.1. Issues of Sample Bias

In general, issues of sample bias revolve around the question, "Were the sample

data collected in a impartial way?" In this case, a large sample was taken (over 70 percent of the population of interest), which reduces the likelihood of bias. Nevertheless, biases can enter the sampling process if those in the sample are systematically different from those excluded from the sample. Both the SET questionnaire and the achievement exam must be scrutinized to ensure that systematic bias has not been introduced into the study.

Sources of any bias introduced by the SET questionnaire are likely to occur if the students who attended the class on the day the survey was administered were somehow different from those who missed class that day or submitted an unusable or incomplete questionnaire. But how might those students be different from those who attended? It seems likely that these two groups may differ in their academic diligence, and this characteristic may be represented in their cumulative grade point average. If this is true, then a comparison of grades of the group in attendance and that group that did not attend may be illuminating.

Table 2 shows the results of a comparison of the cumulative grade point averages of students who took the SET questionnaire versus those who were not present. The table indicates that better students were more likely to participate in the teacher evaluation than lower ranked students. This is not surprising, however, because better students are also assumed to have higher class attendance rates and are more participatory in their education. Moreover, the result that better students are more likely to participate in a teacher evaluation procedure should not contaminate the SET measurement process. It simply means that the SET measurement, *in general*, is provided by the better students in the class.

TABLE 2. SET Participants vs. Non-Participants: A Two Sample T-Test

	PARTICIPANTS	NON-PARTICIPANTS
MEAN GPA	2.72	2.40
VARIANCE	0.46	0.45
OBSERVATIONS	466	163
DF	627	
T-STATISTIC	5.26	
P-VALUE	0.000	

NOTE: The formula assumes equal variances between the populations and a two-tailed hypothesis test.

Sources of bias for the achievement exam are less problematic than the SET because the exam was mandatory in the sections which it was given. Therefore, it can be confidently assumed that the students who took the exam reflect a representative cross-section of introductory economics students.

1.2.2. The Issue of Withdrawals

Another issue in the area of data representativeness deals with the treatment of student withdrawals. This problem is particularly pertinent because a student's willingness to finish the class (also known as "persistence") appears to be highly correlated with both how much the student likes the instructor and how well the student is performing in the class (Douglas and Sulock 1995). One procedure to deal with this problem is to add a "drop" or "no drop" dummy variable to the variables of interest in the specification of the production functions.

The first order of business, however, is a definition of "drop." Surprisingly, this distinction is not as clear as it seems. As a first approximation, a drop may be any student

who officially withdraws from a class according to university records. This definition, however, may be too broad owing to the fact that there are several reasons for not completing a class and this study is interested in measuring the contribution of an individual instructor. Thus, a "teacher-specific" versus "other" reason for dropping the class is desired. Because administrative data are available for the number of credit hours attempted and the number of credit hours completed by each student, a more specific definition of drop can be constructed. In this case, a drop is defined as any student who completed a majority of their attempted hours, but withdrew from their economics class. Therefore, this definition controls for students who dropped a majority or all of their classes, in which case the reason was probably *not* instructor-specific.

Using this definition of student withdrawals, Heckman's (1979) two-step procedure, following Douglas and Sulock (1995) is used. Heckman's first step involves a probit specification to determine whether a student drops or stays based on the right-hand side variables in the production function. This step generates a term called the *inverse Mills ratio* (IMR). Heckman's second step uses the IMR as a regressor in the learning production functions that are estimated. The presence of the IMR in the production functions corrects for any sample selection bias and therefore results in consistent estimated parameters (Douglas and Sulock 1995).

1.3. The Variables and Summary Statistics

Table 3 lists the variables used in this study, and their summary statistics (mean and standard deviation).

TABLE 3. Variable Description and Summary Statistics (571 Observations)

VARIABLE	DESCRIPTION	MEAN	ST. DEV.
	DEPENDENT VARIABLES		
SET	Student Evaluation of Teacher	3.74	0.93
EXAMPCT	Percentage Score on Achievement Test	56.80	14.23
	INDEPENDENT VARIABLES		
ACT	Aptitude Test Score	20.98	3.89
HISCH	High School economics = 1; Other = 0	0.79	0.40
CHOICE	First choice of class = 1; Other = 0	0.87	0.33
EXPGD	Student's Expected Grade	2.64	0.91
GRADE	Student's Actual Class Grade	2.43	1.02
YEAR	Student's year in school	2.17	1.00
PREV	Previous Economics classes = 1; Other = 0	0.35	0.48
MALE	Male = 1; Female = 0	0.54	0.50
GPA	Grade Point Average (from administrative records)	2.63	0.69
BUS	Business Major = 1; Other = 0	0.64	0.48
AGE	Student's self-reported age in years	21.81	4.89
REQ	Required Class = 1; Elective Class = 0	0.79	0.41
ATMHRS	Attempted Hours during semester	14.20	3.14
JOB	Student holds job = 1; Other = 0	0.76	0.43
LARGE	Class 50 or more students = 1; Other = 0	0.12	0.33
NIGHT	Night Class = 1; Other = 0	0.07	0.26
GENMATCH	Instructor and student match gender = 1; Other = 0	0.52	0.50
MICRO	Microeconomics = 1; Macroeconomics = 0	0.30	0.46
LANG	Non-Native English speaker = 1; English = 0	0.24	0.43
TERM	Terminal Degree held by instructor = 1; Other = 0	0.58	0.49
YRSERV	Instructor's years of teaching experience	20.83	13.73

2. Relevant Statistical Estimation Techniques

As discussed in the literature review, there are several statistical practices that have been used in the estimation of educational production functions. Moreover, with the

increased use of computers and the concomitant decline in the real price of computer processing time in the recent past, the development of statistical estimation techniques has flourished. This section discusses several relevant statistical approaches and techniques, along with their advantages and disadvantages. This section concludes with the selection of the statistical technique used to evaluate the data collected for this study.

2.1. Estimating Multiple Outputs

Because *two* educational outputs are being estimated in this study, it is necessary to examine whether, and how, these outputs are interrelated. These interrelationships have both theoretical and empirical implications for the model chosen. Chizmar and Zak (1983, p. 18) have addressed this situation:

> Introducing multiple outputs into the production function raises interesting theoretical and empirical issues. One issue is the manner in which various outputs are interrelated. How one models output interactions in educational production functions may be important. Are the outputs multiple products? If so, are they independently or simultaneously produced? Or are they joint products?[35] Answers to these questions may affect empirical estimates of the educational production function(s).
>
> The relationship among outputs should dictate the model (and estimating technique) one employs to estimate educational production functions.

Chizmar and Zak develop the theoretical and empirical dimensions of the multiple output educational production function. They argue that three situations should be considered. One, if the inputs are shared in the production of the cognitive and affective outputs, then a single equation with multiple outputs should be estimated. This would

---

[35] Multiple products are produced under *separate* production processes, whereas joint products involve the production of more than one output from a *single* production process.

most likely be a "canonical correlation" specification. Second, if the inputs are entirely separate, then two separate output equations should be estimated. And third, if there are feedback relationships between the outputs and inputs a simultaneous equations system is appropriate.

On intuitive grounds, each approach has considerable appeal. Many inputs are involved in the production of educational outputs, and the relative input share is hard to disentangle. However, the empirical literature favors the interpretation that *different* inputs are responsible for the two outputs although sometimes an input is used exclusively in the production of one output (e.g., student choice of class influences their affective output but presumably not their cognitive output) and other times an input contributes to more than one output (e.g., instructor's years of experience influences both affective and cognitive output). When Chizmar and Zak used all three specifications side-by-side, their conclusions were similar with all models.

It is assumed then, for the purposes of this research, that affective and cognitive learning are multiple products and are produced under separate production processes. The appropriate statistical model for this situation requires the specification of separate affective and cognitive equations that are composed of different, but not necessarily entirely so, inputs.

2.2. Panel Data Estimators[36]

The *panel data* statistical technique estimates both the cross-sectional *and*

[36] This section draws heavily from Greene (1993), pp. 464-480.

longitudinal relationships in a data set. Thus, this technique lends itself to use in situations where the data has a natural stratification in both the cross-sectional and longitudinal variables. The cross-sectional variables are commonly economic agents such as households and firms. The longitudinal variable is often a time measurement, but this is not required. For reasons of its newness and statistical power, the panel data technique "is the subject of one of the most active and innovative bodies of literature in econometrics" (Greene 1993, p. 464). Several recent studies have used this technique in economics education (e.g., Bosshardt and Watts 1994; Watts and Bosshardt 1991; Zietz and Cochran 1996).

Because the educational process has a natural stratification between the influence of the teacher and the influence of the student, the panel data technique is well suited to the estimation of these relationships.[37] The application of the panel data technique in this study is more easily understood with an example: assume that many sections of introductory economics constitute the panel. This panel is distinguishable by two features. First, several different instructors are involved, and second, different students and institutional settings are involved. The panel data estimation technique has the advantage of allowing the analyst to disentangle the many attributes of each section. Furthermore, the panel data technique has the advantage of increasing sample sizes. Since all of the instructors are pooled as opposed to being examined separately, more powerful statistical analyses can be wrought from the data.

[37] A useful way to envision panels of data is to imagine similar data sets being "stacked" on one another. In this research, each teacher constitutes one cross-sectional data set.

The major statistical issue in the use of panel data estimators is the *partitioning* of the data into effects of cross-sectional and longitudinal directions. In the case of education, one might ask if the variation in student performance levels is due to *outside* forces (e.g., the instructor), or *inside* forces (e.g., the students). The partitioning of these influences is at the heart of the panel data technique. The distinction between OLS, fixed effects, and random effects models is explained below to clarify the partitioning of panel data sets.

2.2.1. Ordinary Least Squares (OLS) Regression

Ordinary Least Squares regression techniques are widely used in econometric research, and the basic idea of this estimation technique requires no explanation. What does require explanation, however, is the meaning of OLS regression in the context of panel data analysis. In this context, OLS techniques form a baseline against which any "panel effects" can be measured. Stated differently, if there are no significant differences between the data sets that make up the panels, then OLS is a suitable estimator. But if there are panel effects, OLS is an inferior estimator as compared to either the fixed effects or random effects models. For this reason, it is customary to compare the three models side-by-side. If the results suggest that the cross-sectional units are different in some fundamental way, then a comparison of the three models will reject the OLS specification in favor of either the fixed effects or random effects models.

Consider the following model:

(Eq. 1) $$Y_{it} = \alpha_i + \beta' \mathbf{x}_{it} + \varepsilon_{it}$$

where Y_{it} is the performance measure for the t-th student in the i-th instructor's class; α_i is the intercept term; β is a vector of slope parameters; $\mathbf{x}_{it}$ is a matrix of K regressors; and ε_{it} is the disturbance term. The panel consists of n_j students in each teacher's class or classes. If the α_i's are the same across all teachers, it is concluded that teacher effects are unimportant.[38] In this case, it would be concluded that variation in student learning is *not* influenced by the teacher, and is entirely attributable to the students and the institutional setting. This becomes the classical regression model and OLS provides consistent and efficient estimates of α and β. Conversely, if the statistical test identifies that the α_i's are different, then it can be concluded that teachers *do* have an effect on learning outcomes. In this case, either the fixed effects or random effects model is preferred to OLS.

2.2.2. Fixed Effects Estimators

Fixed effects statistical models, also called "least squares dummy variable" models, are operationally the same as identifying each cross-sectional unit in the panel with a dummy variable[39] (Greene 1993). In the case of this study, each teacher is identified with a dummy variable, and the sign and magnitude of the coefficient of the dummy variable indicates whether that particular instructor is different from the average instructor. If the sign of the coefficient is negative, and statistically significant, it can be concluded that the instructor's students, controlled for other factors, performed worse than the average teacher's students. Conversely, if the sign of the coefficient is positive, and statistically

[38] An F-test is used to make this determination.
[39] A "dummy variable," or a "binary variable," has only two values (typically zero and one). This is a common way to estimate the influence of a qualitative characteristic.

significant, it can be concluded that the instructor's students performed better than average. As mentioned earlier, if the dummy variable coefficients are jointly insignificant, it can be concluded that teacher effects are unimportant to output scores. In the words of Greene (1994, p. 469), "The fixed effects model is a reasonable approach when we can be confident that the differences between units can be viewed as a parametric shift of the regression function."

Three practical considerations may limit the use of the fixed effects estimator. First, if there are many data sets (i.e., teachers) compared to observations, then the use of the fixed effects model may suffer statistical problems due to insufficient degrees of freedom. Second, if there is perfect collinearity among any instructor and one of the right-hand side variables, then the use of fixed effects estimators is precluded (e.g., if a particular instructor taught only night classes, and the variable "night class" were used as a regressor). Finally, if the data represent a sample from a larger population and the intercept coefficients (αi's) could be interpreted as random variation instead of instructor-specific variation, then the random effects specification is preferred to the fixed effects model. None of these situations appears to apply in this study. Therefore, if the individual teacher effects are significant, then the fixed effects estimator is the preferred regression technique.

A final technical issue with respect to the fixed effects model is the possible correlation between the right-hand side variables and the teachers. An example may clarify this issue. Suppose that superior students prefer the most experienced instructor because he or she has the best reputation for preparing students for graduate school. Thus, the instructor and at least one of the right-hand side variables (student aptitude) are

correlated. Although some correlation of the data sets is acceptable in a fixed effects specification, too much correlation creates estimation problems. Therefore, a partitioning of the respective influences of the explanatory variables and the instructors is required. This partitioning is done by the random effects model that follows.

2.2.3. Random Effects Estimators[40]

The random effects model is appropriate when only a sample of cross-section units are included in the data set, and the model assumes that the regressors are not correlated with the cross-section effects. The advantage of this specification vis-a-vis the fixed effects model is the additional power of the test due to the increased degrees of freedom and the ability of this model to allocate explanatory power between the cross-section effects and the regressors (Greene 1993). The specification of this model is slightly different from the OLS and fixed effects specifications. Consider the random effects specification below:

(Eq. 2) $$Y_{it} = \alpha + \beta' \mathbf{x}_{it} + \upsilon_i + \varepsilon_{it}$$

There are two differences in this model and the previous specification (Eq. 1). First, the subscript on α has been eliminated. This implies that the intercept term is common to all instructors instead of being an instructor-specific amount. Second, the term υ_i is included. This term is a random disturbance characterizing the ith data set, and is assumed to be uncorrelated with the other right-hand side variables.[41]

---

[40] The random effects model is also called the "variance-components" model.

[41] The inclusion of the υ_i term creates an estimation problem insofar as that term is correlated with the ε_{it} term. This problem can be resolved by estimating the random effects model by the method of *generalized least squares* (see Greene 1993, p.470-472).

The random effects regression technique splits the variation of the left-hand side variable into an instructor-specific effect and an effect attributable to the other right-hand side variables (Greene 1993). As applied to education, this model compares the relative variation within the teachers to the variation within the other right-hand side variables. Three possibilities emerge. First, if there is little or no variation among the teachers and considerable variation among the other right-hand side variables, the classical regression model applies. Second, it there is little or no variation among the right-hand side variables and considerable variation among the teachers, then the fixed effects model applies. Finally, if there is variation among both teachers and the right-hand side variables, the random effects model will allocate this variation among both parties.

Another advantage of the random effects model is its ability to estimate equations when there is perfect collinearity between the teacher and a variable on the right-hand side. As in the example given earlier, if a particular instructor taught only night classes, and "night classes" were a regressor, the random effects model would be able to estimate this equation whereas the fixed effects model would not.

The final pertinent question is: Which model, OLS, fixed effects, or random effects, is the most suitable method to evaluate the data collected for this study? Several considerations come to mind. On purely theoretical grounds, the fixed effects specification is preferred because it seeks to identify an *instructor-specific* contribution to student learning outcomes, and that is much of the focus of this research. Moreover, the data that have been collected fit all of the assumptions of the fixed effects specification.

On the other hand, the random effects estimation procedure is an empirical way to observe the relative variation among teachers and the other right-hand side variables.

Whereas the fixed effects specification *assumes* that the variation in learning outcomes is attributable to instructors, the random effects specification is capable of empirical determination of the relative influence of both teachers and other forces.

To resolve this, an empirical estimation of both models is appropriate. Thereupon, several statistical tests can be invoked to determine which model is more appropriate. Two tests are especially pertinent. First, the F-test mentioned earlier differentiates between the OLS model and the non-OLS models. If there is not a statistically significant difference among instructors, then OLS is preferred to other estimators. Second, the "Hausman Test" can test the fixed effects versus the random effects models. This statistic amounts to a test for *orthogonality*, or independence of the random effects and the regressors. For that reason, it allows the selection between fixed effects and random effects to be made on empirical in addition to theoretical grounds (Greene 1993).

2.3. Multiple Choice Estimators[42]

The final types of estimators to be discussed are the "multiple choice" estimators. These estimators are only appropriate for discrete dependent variables and, therefore, are discussed in the context of the SET score only.

An assumption of the OLS regression model is that the dependent variable is measured on a continuous scale. The achievement test score fulfills this assumption, but the SET score does not.[43] For this reason, many studies in economics education have

[42] These are also known as "polychotomous choice" estimators.
[43] As shown in Appendix 2, the SET score was measured by a ranking system of: outstanding, above average, average, below average, and poor.

used multiple choice estimation techniques (e.g., Brasfield et al. 1993; DeCanio 1986; Mason et al. 1995; Park and Kerr 1995).[44]

Among the popular multiple choice estimators are *ordered probit, ordered logit, multinomial probit,* and *multinomial logit*. As reported in Chapter 2, these estimators have gained popularity in recent years because of the advancement in econometrics and because education variables are often collected in ordinal categories (e.g., A, B, C...and "strongly agree," "agree") as opposed to numbers on a continuous scale. Nevertheless, the majority of the research in economics education continues to use OLS estimation techniques. The question arises: "Is OLS an acceptable substitute for multiple choice estimators?"

Several economists and statisticians have addressed this problem. Two of their responses are recorded below. First, according to Siegfried (1973, p. 71):

> The relevant question is what *practical* difference exists between the empirical results generated by different order preserving measurements of the variables. This question has received some attention. S. Labovitz (1970) shows that for a particular set of ordinal transformations, there is little difference in the statistics derived from the analysis. In another study, R. P. Boyle (1970) concludes "that it will not *usually* be dangerous to assume an interval (cardinal) scale based on categories."...Even though regression analysis cannot be taken literally when applied to ordinal data, it can nevertheless perform an important heuristic and metaphorical function in the interpretation of data.

And according to Labovitz, systems (1970, p. 515): who simulated interval rankings based on twenty randomly-generated, monotonic scoring systems (1970, p. 515):

---

[44] Although much has been made of this issue in the student evaluation literature, the topic does not seem to be discussed in the student grade area, even though it is commonplace to calculate student grade point averages on a cardinal scale when the grade is given on an ordinal scale. The author conducted a literature search and found *no* articles that deal with the impropriety of this transformation.

> Empirical evidence supports the treatment of ordinal variables *as if* they conform to interval scales. Although some small error may accompany the treatment of ordinal variables as interval, this is offset by the use of more powerful, more sensitive, better developed, and more clearly-interpretable statistics with known sampling error...Furthermore, many more manipulations (which may be necessary to the problem in question) are possible with interval measurement, e.g., partial correlation, multivariate correlation and regression, analysis of variance and covariance, and most pictorial presentations.

Clearly, a compelling case can be made for the use of OLS estimators, even if they *technically* violate the assumption of a continuous dependent variable. Moreover, in the case of SET scores collected for this study, there is no reason to suspect that the transformation of the dependent variable from categories to numbers has created any sizable discontinuity or exaggeration of the distance between the ordinal rankings.

3. Estimator Selection

Clearly, there are many theoretical and practical considerations when data are analyzed by statistical models. In the case of this study, a large sample size has been collected which mitigates some statistical problems. Nevertheless, some of the data were collected on an ordinal scale, which make statistical inferences more cumbersome. Questions of causality, simultaneity and interaction effects may also be problematic.

All things considered, a compelling case can be made for the use of panel data techniques. The data have natural strata (the individual instructors) and the nature of this study is to identify the learning contributions attributable to an *individual* instructor. Furthermore, there is variation among students *as well as* instructors, and this variation must be addressed. For these reasons, the panel data estimation technique is used in the statistical analysis that accompanies this study.

Additionally, the connection between the learning equations is tested by a Seemingly Unrelated Regressions (SUR) model. This model treats the two learning equations as a system, and is appropriate when there is a possibility of correlation among the error terms of the models (Pindyck and Rubinfeld 1991). The results of the SUR estimation technique are provided along with a discussion of the appropriateness of the SUR model vis-a-vis the panel data approach.

4. Chapter Summary

This chapter has examined the data and the statistical techniques that are appropriate to this study. After practical and theoretical considerations are made, it is concluded that the data are acceptable and the panel data and SUR estimation techniques are the appropriate statistical procedures to examine the data. Results of the statistical analysis follow.

CHAPTER 5

EMPIRICAL RESULTS

This chapter discusses the empirical results of this study. Included are the actual and adjusted scores of the instructors for both learning measures, the estimated educational production functions, and the *composite* ranking scheme that is constructed from evaluating instructors on more than one learning outcome. The study proceeds in the following sequence. First, the actual SET scores and achievement exam results are reported. This is a useful first approximation to a measure of those instructors who are "reaching" students (the affective score), and those instructors who are "teaching" students (the cognitive score). Second, by the use of estimated production functions, the actual scores are controlled for those variables that may influence the raw scores. Third, several tests are conducted to examine whether there are differences between teachers and differences among students. Finally, a composite ranking scheme is developed using both the cognitive and affective learning measures.

1. Unadjusted Learning Outcomes

The first question to be addressed is an examination of the actual SET and achievement test scores. If these measures are interchangeable, then either one will suffice when measuring teaching effectiveness. On the other hand, if the measures are not interchangeable, then one of the measures must be selected instead of the other, or the measures must be aggregated in some way.

The literature about the interchangeability of SET scores and exam scores has produced interesting findings. One of the popular findings, the oft-cited "Dr. Fox Study," was humorous as well as insightful. Using a professional actor trained to teach in several different styles, this study concluded that high content teaching produces higher cognitive outcomes, highly expressive teaching produces higher student evaluations, and students, in general, were *not* able to discern high content from low content teaching behavior. The "Dr. Fox Effect" was coined for those cases when students confuse academic content with the method of delivery (Williams and Ware 1976).

The Dr. Fox Effect suggests that SET scores and achievement scores are *not* interchangeable. Nevertheless, they may be highly correlated. To test this, one should look at the correlation coefficient between the two measures. If the correlation is high between the actual achievement score and the teacher evaluation, one would conclude that the measures were *practically* interchangeable. This comparison is shown in Table 4. This table reveals an important finding: in this sample there is *no* statistically significant correlation between an instructor's actual SET score and the exam score of their classes ($r = -0.10$). Thus, those instructors that are highly rated by students are not necessarily those instructors that impart high academic content, and vice-versa. Nevertheless, there are some outstanding performers on both outcomes, such as Instructors B and D, and some who perform poorly by both measures (e.g., Instructor K).

Comparisons of this sort may be objectionable to some because they are not controlled for outside factors. For example, what if an instructor taught at an unfavorable time-of-day, or taught students who were ill-prepared for the material? It seems that these actual scores should be controlled for influences outside the control of the instructor. If

TABLE 4. Comparison of Actual SET Scores vs. Actual Exam Scores

INSTRUCTOR	ACTUAL SET SCORE	ACTUAL EXAM SCORE
A	4.29	50.8
B	4.23	56.7
C	4.14	51.9
D	4.10	61.4
G	3.73	55.1
H	3.44	51.1
I	3.28	55.9
J	3.25	71.5
K	3.0	54.4
L	2.67	52.7

NOTE: The Pearson Correlation Coefficient equals -0.10 ($p = 0.78$).

these controls are judged to be important by a statistical standard, then an adjusted ranking scheme could be developed to compare instructors. This line of reasoning anticipates the estimation of the educational production functions that follow.

2. The Estimated Educational Production Functions

This section provides the results of the educational production functions that have been estimated for the purposes of this study. As mentioned earlier, a production function estimates the relationship between an educational output and the several inputs that are involved in the production of learning. Identifying these relationships is essential in the comparison of actual and adjusted outputs. That is, if there is evidence that several inputs contribute to the production of educational outputs, it can be concluded that the raw rankings must be adjusted for those factors. Conversely, if educational inputs other than

the teacher are not important, then *actual* SET and exam scores are adequate measures of those outcomes.

Following the literature, several variables are collected on the students, instructors, and the institutional setting (see Table 3). These variables are used as explanatory variables in the production functions that follow.

2.1. The Cognitive Learning Production Functions

As discussed in Chapter 2, there are several explanatory variables that are commonly used in the specification of cognitive learning production functions. These variables can be characterized as student-specific, teacher-specific, and institutional variables. Consistent with the well-known distinction in economics, these categories can be further sub-divided into fixed and variable inputs.

2.1.1. Student Variables

Fixed input student variables are identified as those variables that are associated with a student's stock of human capital. Those included in the cognitive learning model are: academic aptitude as measured by a popular test (ACT[45]), cumulative grade point average (GPA), year in school (YEAR), high school economics preparation (HISCH), previous college economics preparation (PREV), age (AGE), gender (MALE), subject

[45] The administrative records of some of the students in this study reported the Scholastic Aptitude Test (SAT) as a measure of aptitude. These scores were converted to an ACT score by the commonly-used conversion factor.

interest as represented by academic major (BUS), student-faculty personality match (CHOICE),[46] and whether the course was a requirement (REQ).

Among these variables ACT, GPA, AGE, and MALE have been previously cited as the strongest predictors of cognitive classroom success. Several estimated relationships in the economics field are summarized in Table 5. Among the secondary influences on cognitive learning, the evidence is not compelling with respect to the effect of previous classes in economics at either the college or high school level. Nevertheless, these two variables are included in the models as potential explanatory variables.

Regarding student *variable inputs*, defined as those inputs that can be used more or less intensively, the list includes a measure of competing work commitments (JOB), a measure of competing academic commitments as measured by attempted semester hours (ATMHRS), and a measure of a student's short-term effort in the class as reflected by their expected grade (EXPGD). Somewhat surprisingly, previous findings in educational research have not identified strong relationships between variable student inputs and cognitive learning.

2.1.2. Faculty and Institutional Variables

Similar to the student relationships, faculty variables that are included in the model can also be put into fixed and variable categories. The list of human capital attributes of the faculty includes: terminal degree attained (TERM), number of years of college teaching experience (YRSERV), and non-English native language (LANG).

[46] This variable may also capture a time-of-day influence.

TABLE 5. Student Cognitive Learning Explanatory Variables

	EXPLANATORY VARIABLES					
Author(s)	ACT	GPA	AGE	MALE	PREV ECON	HISCH ECON
Weidenaar and Dodson (1972)	*		*		*	
Tuckman (1975)		*				
Marlin and Niss (1980)	*	*	*			
Watts and Lynch (1989)	*			*		
Watts and Bosshardt (1991)	*			*		
Brasfield, Harrison, and McCoy (1993)	*	*		N	*	*
Gramlich and Greenlee (1993)	*		*	*		
Lopus and Maxwell (1994)	*	*		*		N
Lopus and Maxwell (1995)		*		*		

NOTE: The sign of the estimated coefficient is positive in all cases. An asterisk indicates that the coefficient was significant at the 10% level or better. "N" indicates that the variable was tested, but no statistically significant relationship was found.

The variable input of the faculty cannot be measured directly, and indeed, it is that input that is the focus of this study. If it is shown that the faculty variable inputs are associated with student learning, this will appear in higher than average exam scores for the students of these teachers, ceteris paribus.

The final category of explanatory variables that are included in the model can be loosely called "institutional" variables. These variables reflect the institutional environment in which the classes were conducted, and these institutional conditions are considered to be fixed inputs. Included variables are: the time-of-day the course was offered (NITE), class size over 50 students (LARGE), and microeconomics (MICRO). It is assumed that night classes, because they are three hours long, and large classes, because

of their low faculty-to-student ratio, have a negative impact on student learning. MICRO is an essential control variable because separate micro and macroeconomics tests were used to measure cognitive outcomes.

2.1.3. Estimated Results of the Cognitive Learning Equations

The results of the estimated cognitive production functions are found in Table 6. As shown, the best fitting regression equations use the explanatory variables ACT, GPA, AGE, and MALE. Other variables are shown for comparison purposes.

Because many variables have been mentioned as possible contributors to cognitive learning, variable selection proceeded from a general to a specific model. The first regression included all of the variables listed previously in Sections 2.1.1. and 2.1.2.[47] After the general model was tested, non-contributing variables were eliminated one-by-one using the adjusted R^2 criterion. This procedure works as follows: first, the *weakest* explanatory variable is identified by the value of its low (in absolute value) t-score; second, the variable is eliminated from the model and the regression is re-estimated; and third, the "new" adjusted R^2 is compared with the previous adjusted R^2 and the highest value is selected. This procedure is used to eliminate non-predictive variables, one-by-one, until the model begins to stabilize around a "core" set of significant explanatory variables.

The core set of explanatory variables can be found in Model IV of Table 6. As shown, Table 6 contains the variables ACT, GPA, AGE, MALE, MICRO and IMR (inverse Mills ratio). Several salient points can be made about this core set of variables.

[47] Non-linear forms of all non-dummy variables were tested as well.

TABLE 6. Estimated Cognitive Learning Production Functions

	REGRESSION MODELS and ESTIMATED COEFFICIENTS				
	I	II	III	IV	V
Independent Variables					
CONSTANT	5.08 (0.6)	3.10 (0.4)	7.39 (0.9)	5.40 (0.7)	1.29 (0.1)
ACT	0.98 (5.0)***	1.00 (5.0)***	1.00 (5.0)***	1.00 (5.1)***	1.00 (5.0)***
GPA	4.04 (3.9)***	4.03 (3.8)***	3.63 (3.3)***	3.56 (3.3)***	4.05 (3.8)***
AGE	0.80 (2.8)***	0.93 (3.1)***	0.77 (2.7)***	0.91 (3.0)***	0.94 (3.0)***
MALE	2.75 (2.2)***	2.80 (2.2)***	2.49 (2.0)**	2.52 (2.0)**	2.85 (2.2)***
MICRO	-	-3.90 (-1.1)	-	-4.53 (-1.3)	-4.23 (-1.1)
LARGE	-	-	-	-	1.95 (0.4)
NITE	-	-	-	-	-0.32 (-0.1)
TERM	-	-	-	-	6.28 (1.6)
IMR (Inverse Mills Ratio)	-	-	-5.96 (-1.2)	-6.76 (-1.4)	-
Adequacy Tests					
R^2	0.35	0.36	0.36	0.36	0.36
F (OLS vs. FE)	10.87***	11.12***	10.75***	11.10***	8.72***
Hausman Test	5.94	6.88	5.97	6.96	8.62

NOTES: (1) The dependent variable in these equations is EXAMPCT.
(2) T- values in parenthesis: ** = statistically significant at the 0.05 level, and *** = statistically significant at the 0.01 level.
(3) The F-test compares the Fixed Effects model with the OLS model. A statistically significant coefficient rejects OLS in favor of the FE model.
(4) The Hausman Test compares the Random Effects model with the Fixed Effects model. A large test statistic for the Hausman Test (indicted with *'s) rejects the RE model in favor of the FE model.
(5) The number of observations in all regressions is 299.

First, there is a noticeable absence of instructor effects. The explanation for the lack of instructor effects is straightforward: they are embedded in the error term of the random effects model, and therefore do not show up as specific influences. Stated differently, the rejection of the OLS model in favor of the fixed effects suggests that the panels are different and, therefore, instructor effects *are* important. Each instructor's contribution to cognitive learning can be identified by a fixed effects magnitude.[48] However, the fixed effects model is subsequently rejected in favor of the random effects model, which causes the instructor-specific contribution to be embedded in the random error term. Thus, instructors are indeed different, but the differences cannot be attributed to an instructor individually.

A second feature of the results is the absence of institutional effects. An interpretation of this finding is that time-of-day and class size are not identifiable influences on student learning in this sample, or those influences are captured by the CHOICE variable. Similarly, the MICRO variable was not statistically significant but has a negative coefficient in Models II, IV, and V. An interpretation of this finding is that the microeconomics test may have been slightly more difficult than the macro test, ceteris paribus.

A third notable feature of these results is their agreement with previous findings. The results indicate that the most important variable in cognitive learning appears to be student aptitude. The estimated coefficient of this variable is around 1.0, which means that every one point of aptitude, as measured by the ACT score, contributes to a one

[48] The fixed effects magnitude gives the same result as would a coefficient on an instructor-specific dummy variable.

percent increase on the achievement exam. This finding substantially agrees with earlier authors.

Similar strong and consistent results are found in the GPA variable. For this variable, a one point increase in cumulative grade point average (e.g., from 2.5 to 3.5) is predicted to increase a student's economics test score by about four percent. This finding agrees with a priori expectation: those students with higher grade point averages are generally more academically motivated. Furthermore, this finding agrees with all the authors listed in Table 5.

The AGE and MALE variables also behaved in an expected way. A priori, it is expected that older, more experienced students perform better in economics. In this sample, every year of age contributes about 0.8 percent to the achievement score. The MALE variable is positive and also statistically significant in all models. The magnitude of this effect is about 2.75, meaning that males scored 2.75 percent higher than females, ceteris paribus. An explanation of this finding is beyond the scope of this paper, however, this finding is consistent with the majority opinion in economics education research.

Another variable with important consequences is the IMR (inverse Mills ratio). This variable is at the heart of an important, and controversial, issue in educational research: the sample selection problem. Since students can self-select themselves out of classes by early withdrawal, there is a question of whether the student withdraws because of personal reasons (e.g., illness or work), or because the student does not like the instructor. Clearly, since the focus of this research is on teaching effectiveness, an instructor who has a high drop rate should be identified as ineffective. The IMR makes a statistical determination of whether those who finished the class (and therefore

participated in the SET and the achievement exam) were materially different from those students who did not. In this sample, the insignificant coefficient on the IMR variable indicates that sample selection bias is not significant, and that those students who dropped the class did not do so (necessarily) because of the teacher.

With respect to statistical adequacy tests, the R^2 was uniformly around 36 percent. This figure is within the relative magnitude of previous regression results in this area of educational research. With respect to the appropriate regression model, in all cases the OLS specification is rejected versus the fixed effects model suggesting that instructors, or the different panels, are differentiable. As mentioned before, the fixed effects model is subsequently rejected in favor of the random effects model in all cases, suggesting that there is correlation among the panels (the instructors) and the explanatory variables (the students).

The interpretation of these estimated production functions is one that probably comes as a relief to educators, namely, that teachers *are* differentiable and do have an effect on student cognitive learning. Nevertheless, the finding of Bosshardt and Watts (1994) that student effects seem to be more important than instructor effects appears to be upheld. The positive, and statistically significant, coefficients on the ACT and GPA variables indicate that student human capital is the most influential force in the classroom. In addition, older students seem to learn better than younger students, and males generally perform better than females. All of these findings are consistent with the literature.

2.2. The Affective Learning Production Functions

Like the cognitive production functions, there is a common theme in the specification of affective learning production functions. This theme is again developed along the lines of the student, faculty and institutional input variables. These variables are then divided into fixed and variable categories.

2.2.1. Student Variables

Similar to the cognitive learning equations, much of the work in the affective learning models has been focused on student variables. *Human capital inputs* that students bring to the classroom are their academic interests, their academic abilities and their preparation. For the SET production functions developed in this research, therefore, the following student capital variables are included in the estimation procedure: GPA, ACT, HISCH, YEAR, MALE, GENMATCH, PREV, REQ, BUS, CHOICE and ATMHRS.

Student *variable inputs* are important as well, although they are much more difficult to measure. What can be measured, however, is a student's feeling of "success," and this seems to be another form of short-run labor input. That is, if a student feels successful in the class, that feeling is likely to encourage more effort. The empirical literature on this topic has produced nearly uniform results: the expected grade of a student is convincingly the most important explanatory variable in a student's evaluation of a classroom experience. The findings of research on key student variables are summarized in Table 7.

TABLE 7. Student and Institutional Affective Learning Explanatory Variables

	EXPLANATORY VARIABLES					
Author(s)	EXPGD	TIME OF DAY	CLASS SIZE	REQUIRED CLASS	MAJOR	MALE
Nichols and Soper (1972)	*	*	N			
Kelley (1972)	*			N	*	N
Mirus (1973)	*	N	*	N		
Rose (1975)				N	N	
Dilts and Fatemi (1982)	*		N	N		
Seiver (1983)	*					
Manahan (1983)	*					
Nelson and Lynch (1984)	*				*	*
DeCanio (1986)			N		N	
Mason, Steagall and Fabritius (1995)	*		N			*

NOTE: The sign of the estimated coefficient is positive in all cases. An asterisk indicates that the coefficient was significant at the 10% level or better. "N" indicates that the variable was tested, but no statistically significant relationship was found.

2.2.2. Faculty and Institutional Variables

As was the case in the cognitive learning models, the instructor takes both capital and variable inputs into the classroom. Faculty human capital inputs that have been included in the model are: years of service, terminal degree or lack thereof, gender match with the students, and native language. The research regarding the experience of the instructor in explaining SET scores has been mixed in its explanatory power, but Costin, Greenough, and Menges (1971) report that generally, more experience and higher rank are predictive of higher evaluations. A same-sex match with the instructor has been shown to increase SETs according to Lueck, Endres, and Caplen (1993). Finally, the native

language of the instructor has not been studied regularly as an explanatory variable in SET relationships but it is included as a potentially important control variable.

As for variable inputs, an instructor has the opportunity to work diligently to become a good teacher or put less effort into their teaching performance. The instructor also has the opportunity to be a "hard" or "easy" grader, and thus influence a student's feeling of success in the class. Since this study attempts to measure the variable contribution of the instructor after all other variables have been controlled, it is important to distinguish between good evaluations that redound to hard work, and those good evaluations that result from an easier grading scale. Therefore, the expected grade of the student is an important control variable in the identification of variable faculty inputs.

Finally, the institutional variables in the academic environment must be examined for their influence on the student's evaluation of the teacher. For the sake of classification, these variables are all identified as fixed in nature. The list of appropriate variables includes: the time-of-day the class is offered, the size of the section, and whether the class is microeconomics or macroeconomics. These variables are represented by NITE, LARGE, and MICRO, respectively. Some previously estimated relationships for the time-of-day and class size variables are identified in Table 7.

2.2.3. Estimated Results of the Affective Learning Equations

The results of several estimated SET equations are shown in Table 8. These equations identify those variables that best explain an instructor's student evaluation score. The results agree, for the most part, with previous research and a priori expectation. Because many variables have been mentioned as possible contributors to

a teacher's student evaluation score, variable selection again proceeded from a general to a specific model. The early regressions included all of the student, faculty and institutional variables listed in the previous sections.[49] After this general model was estimated, non-contributing variables were eliminated using the adjusted R^2 criterion. The best fitting regressions, by this criterion, are listed in Table 8.

As anticipated, a significant influence in a teacher's evaluation is explained by the student's expected grade. The coefficient on the EXPGD variable is in the magnitude of 0.33, which suggests that each letter grade improvement brings about a one-third point increase in the student's evaluation of the instructor. This finding is important for another reason: this variable, unlike the other variables, can be manipulated by the instructor.

The second variable that was statistically significant throughout all equations was the CHOICE variable. This finding suggests that there is a positive association between a student's choice of instructor and their rating of that instructor, and those students that took the class of their first choice gave the instructor a higher evaluation by about 0.42 points. This finding is consistent with a priori expectations, and is important insofar as this variable is *not* commonly included in the specification of SET production functions, presumably because it is a hard datum to collect. Instead, the variable most often used to model student choice is a time-of-day variable. There is a subtle, and important, difference between these two specifications. The time-of-day variable *assumes* that there are preferred times to take the class. While this is no

[49] Non-linear forms of all non-dummy variables were tested as well.

TABLE 8. Estimated Affective Learning Production Functions

	REGRESSION MODELS and ESTIMATED COEFFICIENTS				
	I	II	III	IV	V
Independent Variables					
CONSTANT	3.01 (10.4)***	2.78 (11.4)***	3.07 (10.5)***	2.81 (7.0)***	2.89 (6.3)***
EXPGD	0.33 (6.1)***	0.33 (6.0)***	0.35 (6.3)***	0.33 (6.1)***	0.33 (6.0)***
CHOICE	0.42 (3.4)***	0.43 (3.5)***	0.42 (3.4)***	0.42 (3.5)***	0.42 (3.4)***
GENMATCH	0.22 (2.6)***	0.17 (2.0)**	0.19 (2.3)***	0.24 (2.7)***	0.15 (1.5)
LANG	-0.57 (-2.4)***	-0.58 (-2.5)***	-0.58 (-2.5)***	-0.56 (-2.4)***	-0.59 (-2.5)***
BUS	0.17 (1.8)*	0.20 (2.2)***	0.17 (1.9)*	0.24 (1.8)*	0.16 (1.0)
GPA	-	-0.17 (-2.5)***	-0.11 (-1.4)	-	-0.18 (-2.1)***
ACT	-0.03 (-2.7)***	-	-0.02 (-1.8)*	-0.03 (-2.5)***	-
IMR (Inverse Mills Ratio)	-	-	-	0.69 (0.7)	-0.34 (-0.3)
Adequacy Tests					
R^2	0.36	0.36	0.36	0.36	0.36
F (OLS vs. FE)	6.17***	6.09***	6.03***	6.19***	6.04***
Hausman Test	5.81	6.78	7.16	5.83	6.86

NOTES: (1) The dependent variable in these equations is SET.
(2) T- values in parenthesis: * = statistically significant at the 0.10 level, ** = statistically significant at the 0.05 level, and *** = statistically significant at the 0.01 level.
(3) The F-test compares the Fixed Effects model with the OLS model. A statistically significant coefficient rejects OLS in favor of the FE model.
(4) The Hausman Test compares the Random Effects model with the Fixed Effects model. A large test statistic for the Hausman Test (indicated with *'s) rejects the RE model in favor of the FE model.
(5) The number of observations in all regressions is 344.

doubt partially true, a more important distinction is whether the student *chose* the time they took the class, *even* when the class was offered at an "inferior" time-of-day.

The third variable that was significant throughout all the equations was the language variable (LANG). Moreover, the coefficient on this variable is negative. An interpretation of this finding is that student's *penalize* their instructors about 0.50 points on the student evaluation in cases where the instructor's native language is *not* English. Interestingly, in these estimated regressions, the language variable is the only faculty variable that is statistically significant.[50]

The fourth variable that is generally significant is the faculty-student gender match (GENMATCH). Students in this sample rate their instructor higher, by about 0.20 points, if both student and instructor were of the same gender. This finding agrees with Lueck, Endres, and Caplen (1993), and has interesting ramifications. Since the economics faculties in colleges and universities are typically dominated by men, and because more women appear to be selecting the business major as opposed to more "traditional" fields, this finding suggests that there is less gender-matching in schools of business than there has been historically. This may lower student evaluations in "male-dominated" fields as a consequence.

The fifth notable relationship was among the two measures of student academic prowess, ACT and GPA. Each was significant in several equations. The signs of the coefficients, however, are negative for both variables. The interpretation of this finding is unclear but, as a reasonable explanation, this may suggest that higher aptitude and higher

[50] Again, recall that some faculty effects are captured in the cross-sectional panels.

performing students are more critical of their instructors, ceteris paribus. An alternate explanation is that *all* students reflect their expected grades in some reward for the teacher, but the better students do it less than other students. In any event, the magnitude of the estimated coefficient is small for each variable, so the net effect of this variable on the instructor's SET score is small.

The final variable with predictive power in these equations is BUS, or the dummy variable that identifies College of Business majors.[51] The positive coefficient on this variable indicates that business majors, in general, give slightly more favorable teacher evaluations than non-business majors. This finding is intuitively appealing because business students probably make a positive association with their core classes. Otherwise, it seems, they would not select a major in business. Moreover, this finding agrees with the dominant position in the literature that majors tend to rate their instructors higher than do non-majors.

It is also noteworthy that the IMR variable was insignificant in the SET regressions as it had been in the EXAM regressions. This suggests that possible bias from students who drop the class is not an acute problem in this sample.

With respect to statistical adequacy tests, the SET equations produce results very similar to the EXAM equations. Again, R^2 is uniformly around 36 percent. With respect the appropriate regression model, in all cases the OLS specification is rejected versus the fixed effects model suggesting that instructors are differentiable. Furthermore, the fixed effects model is subsequently rejected in favor of the random effects model in all cases.

---

[51] Because there are relatively few economics majors in the College of Business vis-a-vis the other fields, it was deemed to be more useful to identify the general area of study rather than the major.

This suggests that there is correlation among the panels (the instructors) and the explanatory variables (the students).

3. A Seemingly Unrelated Regressions Approach

A final topic in the statistical area is relevant before this study turns to the implementation of the preceding regression results, and that is a discussion of the connection between the two estimated learning equations. Because an individual student represents a single observation in both of the learning equations, it is possible that the error terms on the equations are correlated. Improved estimates may be possible by setting up a system of equations and simultaneously estimating the EXAM and SET relationships. A by-product of this estimation technique is that it may allow some of the teaching and institutional effects to surface that were *embedded* in the error term of the random effects model.[52]

An appealing specification in this situation is the *seemingly unrelated regression* model (SUR). This is an econometric model that simultaneously estimates two or more related equations (Pindyck and Rubinfeld 1991). Since the exam and SET scores bear a close resemblance to each other, SUR may provide additional insight into these learning equations. The results of the SUR estimations are shown in Table 9. They are materially the same as the panel data estimations, but several additional variables emerge as statistically significant. This is not surprising, however, because the effect of the instructor was borne by the cross-section term in the panel data model, and in the SUR model the

[52] Recall that the fixed effects model explicitly identifies the contribution of *each* instructor. The random effects model, on the other hand, embeds the effect of the instructor in the random error term.

TABLE 9. Estimated SUR Production Functions

	REGRESSION MODELS and ESTIMATED COEFFICIENTS	
Independent Variables	**EXAMPCT**	**SET**
CONSTANT	11.96 (1.2)	2.89 (8.4)***
ACT	1.01 (4.2)***	-0.03 (-2.1)***
CHOICE	3.53 (1.7)**	0.30 (2.1)***
GPA	3.26 (2.5)***	-
AGE	0.52 (1.5)	-
EXPGD	-	0.44 (7.0)***
TERM	4.33 (2.8)***	-
LARGE	-	-0.45 (-3.0)***
HISCH	-4.04 (-1.9)**	-
LANG	-	-0.71 (-6.3)***
GENMATCH	-	0.17 (1.8)**
YRSSERV	-	0.01 (1.6)**
Adequacy Tests		
R^2	0.20	0.32
F	9.22***	14.17***

NOTES: (1) T- values in parenthesis: ** = statistically significant at the 0.05 level, and *** = statistically significant at the 0.01 level.
(2) The number of observations in both regressions is 244.

data are *not* partitioned by faculty member. The procedure for selecting the appropriate variables proceeded the same as for the panel data estimates. That is, the first regression was estimated using all the variables in both equations. Insignificant variables were subsequently eliminated from each equation based on their marginal contribution to adjusted R^2. This procedure was continued until a core set of independent variables was found.

As before, the EXAM equation was dominated by the ACT and GPA variables, but some of the previous explanatory variables were no longer significant and some new variables entered the equation. Two previously significant variables, AGE and MALE, lost their explanatory power under the SUR specification. The AGE variable was positive, but no longer statistically significant ($t = 1.5$), and similarly, the MALE variable was not predictive in the SUR model. Instead, several new variables entered the equation. Both the terminal degree and student choice variables entered the equation with positive coefficients, and the high school economics variable entered the model with a negative coefficient. All of these relationships follow from logic or previous research. Specifically, the effect of the instructor holding a terminal degree produced a 4.33 percent improvement in achievement scores. Similarly, student choice of instructor seems to contribute to higher exam scores. This has some intuitive appeal because it seems likely that a student would select an instructor who the student thinks will match his or her learning style. Finally, having taken a class in high school economics appears to have deleterious effects on college-level economics achievement scores. While this finding seems counter-intuitive, it does agree with previous studies in economics education (e.g., Highsmith and Baumol 1991; Lopus and Maxwell 1994).

For the SET equations, the core of variables (EXPGD, CHOICE, GENMATCH, LANG, ACT) is the same under the SUR specification as under the panel data approach, however two new variables enter the equation. The large class variable enters the equation with a negative coefficient, and the years of service variable enters with a positive sign. The addition of these new variables is largely attributable, as it was for the EXAM equation, to the fact that the data are not partitioned by instructor as they were for the panel specification. It should also be mentioned that both of the additional variables to the SET equations are consistent with other research in this area (DeCanio 1986, and Seigfried and Fels 1979, respectively).

A comparison of the SUR and panel data approaches may be instructive. As mentioned earlier, the results of the two approaches are materially the same. Nevertheless, the subtle differences that emerge deserve mention. The panel data technique produces the higher explanatory power, with the R^2 values in the magnitude of 0.36, whereas the R^2 values in the SUR specification are somewhat lower. Furthermore, the panel data approach appears to capture the subtle effect of a teacher's "personality" that is not identified as a separate regressor. Specifically, the fact that the OLS specification is rejected in favor of the fixed effects model indicates that teachers are different and some intangible, "personality effect," may be at work.

Even though they are not used in the formulation of the composite index, the SUR estimates make a contribution to this study. In particular, they reveal some of the instructor effects that are embedded in the panel data model. For example, the significance of the CHOICE variable in the cognitive learning equation suggests that

students may choose instructors for learning reasons in addition to reasons of convenience. Similarly, the significance of the terminal degree variable (TERM), suggests that having a more highly educated faculty appears to contribute positively to student cognitive learning.

The same subtle influences of the SUR estimates are apparent in the SET equations, and the loss of R^2 is negligible vis-a-vis the panel data estimates. In the SUR case the institutional variable measuring class size is significant, supporting the widely-held notion that students do not like "big classes." Also significant is the effect of the instructor's years of service. This effect, although small, suggests that students give higher ratings to more experienced teachers, ceteris paribus.

Based on the goodness-of-fit criterion, this study uses the panel data estimates in the measurement of teaching effectiveness that follows.

4. Differences in Classes and Instructors

Considering the findings from the regression equations, a strong case can be made that human capital and the behavior of students and faculty have important influences on both cognitive and affective learning outcomes. In other words, *actual* achievement exams and student evaluations appear to be biased and incomplete measures of student outcomes. Adjusting the actual scores for the extraneous influences identified by the regression equations will mitigate the bias and produce better measures of teaching effectiveness.

The next condition in developing a ranking scheme is the determination of whether or not the classes vary in material ways. Stated differently, what if all classes had the

TABLE 10. High Aptitude Classes vs. Low Aptitude Classes: A Two Sample T-Test

	HIGH APTITUDE CLASSES	LOW APTITUDE CLASSES
MEAN ACT	21.86	20.03
VARIANCE	14.51	14.60
OBSERVATIONS	102	115
DF	215	
T-STATISTIC	3.52	
P-VALUE	0.00	

NOTE: The T-test formula assumes equal variances between the populations and a one-tailed hypothesis test.

same percentage of high-aptitude students? And what if an individual student feels more successful because he or she expects a good grade in a particular class, but all instructors exhibit identical grading behaviors? In these situations there would be no way to distinguish good evaluation and exam scores from bad scores. Therefore, it is necessary to identify whether classes are different in material ways. Because each production function has one variable that is a particularly strong predictor of the dependent variable (ACT for the cognitive outcomes and EXPGD for the SET outcomes), it seems prudent to look for differences among classes with respect to the aptitude of the students, and differences in teachers with respect to their grading behaviors.

4.1. Class Aptitude

Table 10 identifies high aptitude and low aptitude classes. To make the distinction between high and low aptitude, an average ACT score was calculated for all classes. Then, the high aptitude classes were defined as the three highest sections in the sample,

and the three lowest sections were defined as low aptitude classes. The difference in the section average between the high and the low classes is 1.83 points on the ACT scale, and this difference is statistically significant at the 0.01 level. Although most classes are more similar than different in terms of the aptitude of their students, this example shows that a particular instructor can randomly be assigned to a low or high aptitude class. Furthermore, this finding suggests that some sections are expected to perform more poorly on the achievement exam than others, ceteris paribus.

4.2. Faculty Grading Behaviors

It is also instructive to see if instructors vary systematically with respect to their grading behaviors, or if their grading behaviors are largely related to the quality of the students they teach. Using logic similar to the previous section, the sample of instructors was partitioned so that the bottom-three low graders could be compared with the top-three high graders. As shown in Table 11, the students of the "hard" graders had a mean grade point average of 2.15, whereas the mean grade point average for the students of the three "easy" graders was 2.92. These results are consistent with a hypothesis that there are significantly different grading behaviors within the group of sampled instructors, and a case can be made that there are some easy graders and others with harder grading standards. This finding strongly suggests that raw SET scores must be adjusted to account for the variability in grading standards. It should be mentioned that the preceding comparison involves a student's *actual grade*, not their *expected grade*. Because students do not know their actual grade when the teacher's evaluation is conducted, the expected grade is just a student's estimation. Not surprisingly, there is a high correlation between

TABLE 11. "Hard" Graders vs. "Easy" Graders: A Two Sample T-Test

	"HARD" GRADERS	"EASY" GRADERS
MEAN ACTUAL GRADE	2.15	2.92
VARIANCE	0.90	0.90
OBSERVATIONS	128	118
DF	242	
T-STATISTIC	-6.34	
P-VALUE	0.00	

NOTE: The T-test formula assumes equal variances between the populations and a one-tailed hypothesis test.

the actual and expected grades (0.69 in this sample), and it seems safe to assume that the grading behavior of the instructor is understood by the students at the time the student evaluation is done.

5. The Prediction Equations

The previous sections have established that students and faculty differ in important and measurable ways. The next step in the evaluation of teaching effectiveness is to adjust the actual scores on the achievement exams and SETs for extraneous influences. The first step in this process is the creation of *prediction equations.* The prediction equation is defined as that equation that best uses the estimated regression coefficients in conjunction with the instructor-specific input variables to predict both an exam and a SET score for each teacher.[53] Using this logic, the predicted value of each learning outcome is based

[53] Note that the instructor has only *one* predicted score for each learning output, even if the instructor teaches more than one section. This seems consistent with the way most administrators evaluate the faculty. A class-by-class prediction equation would be a straightforward extension of this idea.

upon the educational inputs and their relative influence as measured by the regression coefficients. Thus, the prediction equation best typifies each instructor's expected score.

The panel data regression results in Tables 6 and 8 reported several plausible learning equations. One equation each must be chosen from each table, and some judgment must be exercised. For the sake of this study, the equation listed as Model I in each table is chosen. The rationale for this selection is that the most *parsimonious*, or uncomplicated, model is chosen to simplify the calculations and interpretations. The prediction equations, in numerical form, are provided below:

Predicted Exam Score = 5.08+0.98 (ACT) +4.40 (GPA)+0.80 (AGE) +2.75 (%MALE);

Predicted SET Score = 3.01+0.33 (EXPGD) +0.42 (%CHOICE) +0.22 (%GENMATCH) +0.57 (LANG) +0.17 (%BUS) -0.03 (ACT)

From these prediction equations, specific values are calculated on an instructor-by-instructor basis for both the achievement exam and the student evaluation of the teacher. The results of these calculations are found in Table 12.

5.1. Using the Prediction Equations to Measure Teaching Effectiveness

The next step in the process is to evaluate the predicted EXAM and SET scores versus the actual performance. This comparison yields two important findings: one, the difference (if any) between what could reasonably be expected from an instructor and what was actually observed; and two, an ordinal ranking based on the percentage deviation between the actual and expected outcome after adjusting for extraneous influences. These comparisons are found in Tables 12 and 13, respectively.

TABLE 12. Comparison of Raw Exam Scores vs. Predicted Exam Scores

INSTRUCTOR	ACTUAL EXAM SCORE	PREDICTED EXAM SCORE	PERCENTAGE DIFFERENCE	RANK OF % DIFFERENCE
A	50.8	56.0	-10.2	10
B	56.7	56.4	0	4
C	51.9	55.2	-6.4	9
D	61.4	55.3	9.9	2
G	55.1	56.4	-2.4	6
H	51.1	53.8	-5.3	8
I	55.9	54.6	2.3	3
J	71.5	56.3	21.2	1
K	54.4	54.7	0	4
L	52.7	54.9	-4.2	7

NOTE: The assignment of alphabetic labels follows the pattern of the "Research SET Rank" in Table 1.

From Table 12, entitled "Comparison of Raw Exam Scores vs. Predicted Exam Scores," several important observations can be made. First, the actual and predicted scores are dissimilar insofar as the actual scores show a larger spread than the predicted scores. The range of the actual scores is 20.7 percentage points (71.5-50.8), whereas the range of the predicted scores is only 2.6 percent. As a first approximation, this finding suggests that actual learning has a higher variance than expected learning, and the *teacher seems to make a difference.* Second, the "Rank of Percentage Difference" column seems not to be highly correlated with the "Instructor" column. Since the instructors were placed in rank order by raw SET score in this table, this finding is suggests the SET ranking is not good predictor of cognitive learning outcomes in this sample.

In Table 13, the same logic is applied as in Table 12 but in this case SET scores are compared. As was the case for the exam scores, there is more variability in the actual

TABLE 13. Comparison of Raw SET Scores vs. Predicted SET Scores

INSTRUCTOR	ACTUAL SET SCORE	PREDICTED SET SCORE	PERCENTAGE DIFFERENCE	RANK OF % DIFFERENCE
A	4.29	4.07	5.4	3
B	4.23	3.81	10.7	1
C	4.14	3.93	5.1	4.5
D	4.10	3.95	3.7	6
E	3.97	3.78	5.1	4.5
F	3.9	4.14	-5.8	10
G	3.73	3.91	-4.6	8
H	3.44	3.25	5.9	2
I	3.28	3.66	-10.5	11
J	3.25	3.21	1.2	7
K	3.0	3.16	-5.0	9
L	2.67	3.62	-26.2	12

NOTE: The assignment of alphabetic labels follows the pattern of the "Research SET Rank" in Table 1.

SET scores than the predicted scores, but the differences are not as extreme. Namely, the range from the high to low actual score is 1.62 (2.67 to 4.29), whereas the range from the high to low predicted score is 0.98 (3.16 to 4.14). The interpretation of this is like before: *teachers seem to matter* as measured by their student evaluation scores. There is, however, a positive correlation between the "Instructor" column and the "Rank of % Difference" column (Spearman Rank Correlation Coefficient = 0.72, p value less than 0.01). This suggests that the rank order changes when an adjustment for extrinsic forces is made, but this change is not large. This should not be greatly surprising, however, because grading behavior is embedded in both the predicted and actual scores and, as shown earlier, grading behavior within the department is non-uniform.

5.2. A Comparison of Cardinal and Ordinal Teaching Effectiveness Measures

The prediction equations were developed to control educational learning measures for influences outside the control of the instructor. The equations have also provided a way to compare instructors on both cardinal and ordinal scales. Furthermore, the prediction equations have shown that *controlled* measurements of teaching effectiveness are different from *uncontrolled* measurements. Considering these findings, the evidence suggests that extrinsic forces should be accounted for when evaluating teaching performance.

This section proposes a way to implement the results of this research. First, an examination of the relationship between the cardinal and ordinal measures is contemplated. Because instructors have traditionally been ranked against their peers, the ordinal results are the point of departure. For example, one might be interested in the number of instructors who have changed relative position after extrinsic influences have been considered. This analysis is done for both of the learning outcomes.

An examination of Table 14, "Comparison of Actual vs. Predicted SET Ranks," shows that using the prediction equations causes a re-ranking of the faculty in this sample. The general ranking of the instructors after adjustment is similar to that before adjustment, as evidenced by the 0.76 rank correlation between the "Actual" and "Predicted" columns. Most instructors only changed a position or two after extrinsic forces were accounted for, and that could easily be attributed to random chance. Three professors, however, are more than three ranks from their expected performance. This change of ranking is, indeed, reflective of abnormally good or bad performance. For example, Instructor "B" performed much better than predicted, as evidenced by an actual rank of second versus a

TABLE 14. Comparison of Actual vs. Predicted SET Ranks

INSTRUCTOR	ACTUAL SET RANK	PREDICTED SET RANK
A	1	2
B	2	6
C	3	4
D	4	3
E	5	7
F	6	1
G	7	5
H	8	10
I	9	8
J	10	11
K	11	12
L	12	9

NOTE: The Spearman Rank Correlation Coefficient equals 0.76 ($p = 0.00$).

predicted rank of sixth, or a difference of four ranking positions. On the other hand, Instructors "F" and "L" performed much worse than was expected. Instructor "F" was the most deficient, falling five positions from the rank predicted, and Instructor "L" fell three ranks. Both of these instructors would be identified as deficient performers, but importantly, Instructor "F" was in the middle of the group when measured by raw ranks and would not have been considered deficient by the traditional standard.

Turning to the exam ranks in Table 15, the same analysis is applied to the achievement test scores. In this case, the "Instructor" column is ranked by the actual scores that the students made on their achievement exams. The "Predicted Exam Rank" column uses the prediction equation to estimate a rank based on the conditions peculiar to each instructor's sections. In this case, however, the results are quite different from the

TABLE 15. Comparison of Actual Exam Ranks vs. Predicted Exam Ranks

INSTRUCTOR	ACTUAL EXAM RANK	PREDICTED EXAM RANK
J	1	3
D	2	5
B	3	1.5
I	4	9
G	5	1.5
K	6	8
L	7	7
C	8	6
H	9	10
A	10	4

NOTE: The Spearman Rank Correlation Coefficient equals 0.41 ($p = 0.12$).

SET results. A cursory look at Table 15 indicates there is a low correlation between the actual and predicted ranks. Three observations come to mind. First, there is a low relationship between actual exam scores and actual SET scores. This can be observed by noticing how "out of alphabetical order" the "Instructor" column is. Second, the actual scores *do not* closely reflect what was predicted, as is the case for the evaluation scores. This indicates, ceteris paribus, that the correction of raw scores for extrinsic influences is particularly important when comparing exam scores, and much more important than for SET scores. Finally, considering exceptional producers of cognitive learning, four instructors stand apart from the group. Those that performed significantly better than expected were Instructors "I" and "D," whose actual performance was higher than predicted by four and three ranks, respectively. As for deficient performance, Instructors "A" and "G" performed six and three and one-half ranking positions, respectively, below

what could reasonably be expected. The rest of the sample was within two positions from their expected ranks, and variation of that magnitude might be attributed to chance.

It is notable that on the basis of actual exam scores, both Instructors "D" and "A" would have been correctly identified as different from the average. On the other hand, Instructor "I" would not have been recognized for superior performance, and Instructor "G" would not have been identified as deficient.

As a final comparison, the cardinal values instead of the ordinal ranks are used. Using these numbers, a system could be envisioned that sets a *benchmark* against which relative performance could be measured. Although this benchmark could take on several possible values, assume that a ten percent deviation from the predicted score would indicate *exceptional* performance. Who would be rewarded, and who would be penalized under this scheme? For the exam scores (see Table 12), Instructors "J" and "D" would be rewarded, whereas "A" would be sanctioned. For the SET scores (see Table 13), "B" would be rewarded, whereas "I" and "L" would be sanctioned. Under this scheme, three professors are identified as "meritorious," three are identified as "non-meritorious," and the remaining six would be considered typical instructors. It is worthy of note that *none* of the exceptional instructors were recognized in *both* output categories. This reflects the imprecise relationship between exam scores and teaching evaluations. Furthermore, this thinking anticipates the creation of the *composite* index that follows.

6. A Composite Measure of Teaching Effectiveness

Most administrators would agree that both SET scores and achievement exams are meaningful measures of teaching effectiveness. Furthermore, much evidence has been

TABLE 16. The Composite Ranking Scheme I

INSTRUCTOR	ADJUSTED SET RANK	ADJUSTED EXAM RANK	COMPOSITE SCORE I	COMPOSITE RANK I
A	3	10	6.5	5.5
B	1	4	2.5	1
C	4.5	9	6.8	7
D	6	2	4	2.5
G	8	6	7	8.5
H	2	8	5	4
I	11	3	7	8.5
J	7	1	4	2.5
K	9	4	6.5	5.5
L	12	7	9.5	10

NOTE: The SET ranks contain more instructors than the Exam rankings, which accounts for the discontinuity in the alphabetic labels in the "Instructor" column and the values higher than 10 in the "Adjusted SET Rank" column.

offered to indicate that the two measures are *not* the same. This, and other, research has shown that raw scores must be adjusted to reflect conditions outside the control of the teacher. The final stage in this study develops an evaluation system that merges both adjusted output measures into a single, *composite*, measure. This is done in two ways, first by ranks and second by deviations from the predicted score.

The ordinal ranking of instructors takes the unadjusted SET and exam scores, adjusts them for extraneous influences, and calculates a simple average of the ranks to construct a composite score (Composite Score I). This score is then rank ordered, and that is the basis for teacher comparisons. These calculations are shown in Table 16.

As shown, the "Composite Rank I" column is materially different from the "Instructor" column, which indicates that the measurement of teaching effectiveness is

considerably different under this scheme than under the traditional student evaluation scheme. Whereas many teachers hold their relative position (e.g., Instructors "B" and "L"), several other significant re-ranks occur (e.g., Instructors "A" and "J"). If it is believed that considering both output measures is preferred to a single measure of teaching effectiveness, then this scheme provides a better way to make that measurement.

A second way to develop a composite scheme relies on an inter-instructor comparison using the deviations from the predicted score. As in the previous scheme, each instructor's predicted exam and SET scores are compared with their actual scores, and the sum of these gives a total deviation. A simple average for each instructor is then computed from these deviations. Finally, instructors are ranked according to this average deviation. The results of this ranking scheme can be seen in Table 17. Interestingly, the results of this table are quite similar to the results of Table 16 (the rank correlation coefficient between the two is 0.92), suggesting that both ways of evaluating faculty will produce similar results. The outstanding instructors in Table 15 ("B," "D," and "J") were also ranked in the top-three on the percentage deviation grounds. Similarly, the bottom three performers in Table 16 ("G," "I," and "L") were also the lowest three performers in Table 17. These findings suggest that either of the two composite ranking schemes produce consistent estimations of teaching effectiveness. Furthermore, because the composite ranking scheme values *both* educational outputs, a teacher can demonstrate effectiveness over a wider range of outputs, and this results in fewer incidents of reward or discipline than either of the standards taken separately.

TABLE 17. The Composite Ranking Scheme II

INSTRUCTOR	% DIFFERENCE SET SCORE	% DIFFERENCE EXAM SCORE	COMPOSITE SCORE II	COMPOSITE RANK II
A	5.4	-10.2	-2.4	6
B	10.7	0	5.3	3
C	5.1	-6.4	-0.7	5
D	3.7	9.9	6.8	2
G	-4.6	-2.4	-3.5	8
H	5.9	-5.3	0.3	4
I	-10.5	2.2	-4.2	9
J	1.2	21.2	11.2	1
K	-5.0	0	-2.5	7
L	-26.2	-4.1	-15.3	10

NOTE: The SET ranks contain more instructors than the Exam rankings, which accounts for the discontinuity in the alphabetic labels and the values higher than 10 in the "Adjusted SET Rank" column.

7. Discussion

It has been argued throughout this paper that student evaluations of teaching are incomplete measures of teaching effectiveness. The composite ranking scheme has proposed a way to deal with the two major sources of incompleteness in the student evaluation process: evaluations are not adjusted for extraneous forces in the classroom, and they do not measure cognitive learning. The composite measure, therefore, is advanced as a preferred way to measure teaching effectiveness.

The movement to standardize and measure teaching effectiveness has been around for several years. Moreover, there are several noticeable efforts in this direction that have been undertaken by schools in recent years. Standardized departmental exams and departmental adoption of common textbooks stand out as examples of this movement.

The movement to implement a common set of academic standards appears to be underway. A common way of measuring outcomes, such as the composite measure proposed in this paper, seems to be one way to achieve more standardization.

With respect to implementation of the composite measure, several small steps must be taken. First, *some* standardized exam must be administered in all classes that have multiple sections. Second, more detailed information on student and institutional attributes must be secured to adjust achievement and SET scores for extrinsic influences. Finally, reward schemes must be developed to enforce the motivational consequences of the measurement scheme. If these steps are taken, the composite measurement scheme could become operational.

8. Chapter Summary

This chapter has reviewed the several numerical measures of teaching effectiveness. As a first approximation of teaching effectiveness, the traditional standard uses an inter-instructor comparison of student evaluations. It has been argued that this technique is deficient for three reasons: cognitive learning is not included in this measure; extrinsic influences may be important; and instructors may lower their grading standard in an attempt to *buy* good evaluations.

The chapter proceeded by producing evidence that cognitive learning is *not* adequately reflected in student evaluations. Next, evidence was presented that showed that extrinsic influences *are* important, and therefore must be controlled when making inter-instructor comparisons. Finally, evidence was adduced that instructors in this sample

have individual grading standards, and the grades that students receive are not necessarily commensurate with student learning.

Considering this evidence, a final step is to merge the two learning outcomes, appropriately adjusted for extrinsic influences, into a composite measure of teaching effectiveness. It is shown that whether this composite index is constructed using ranked or cardinal data, the results are similar in identifying effective or ineffective instruction.

CHAPTER 6

CONCLUSIONS

This study has examined several issues involved in measuring teaching effectiveness and in the identification of exceptional teaching performance. Economic theory was presented to show why performance measurement is important in the resource allocation, motivation, and reward schemes that apply to higher education. It was also noted that the existence of a measurement scheme presents several problems because economic agents often change their behavior under observation. Teaching performance was measured using two methods of assessment. Finally, an index was developed that purports to measure a teacher's overall effectiveness. The conclusions of this study follow.

1. Findings

Some findings of this study can be considered corroboration of earlier research, and others of which can be considered to be new additions to the knowledge base of economics education. These findings are organized as performance measurement findings, SET findings, and achievement exam findings.

1.1. Performance Measurement Findings

The findings with respect to the performance measurement of the instructors are of primary interest in this study. First among these findings, it was shown that there is a low correlation between the raw exam scores and the raw SET scores. This implies that these scores are not interchangeable measures of teaching effectiveness for the instructors in this

sample. Next, it was found that control variables are essential when teaching effectiveness is measured. When the raw scores were controlled for extrinsic influences, a *baseline* performance level was established. From that baseline, exemplary performers could be distinguished from deficient performers. Also, it was also shown that, once a comparison to a baseline was done, performance could be measured by either cardinal deviations from the baseline or ordinal ranks.

Finally, it was shown that some good performers on an unadjusted scale would be judged deficient on an adjusted scale. Conversely, it was shown that some low performers on a raw scale would be identified as above-average performers on an adjusted scale. These findings resolve some of the shortcomings of the traditional measurements of faculty performance.

1.2. SET Findings

The consensus across educational research is that student evaluations of their teachers are strongly related to a student's expected grade. That finding was confirmed in this sample, and this suggests that students "like" their teacher in proportion to how well the students think they are doing in the class. While not surprising, this finding also suggests that teachers can *buy* good evaluations with an easy grading policy. Therefore, it is recommended that SETs be evaluated in the context of the instructor's grading behavior.

Other relationships have been discovered as well. It has been shown that student choice of teacher plays a role in how well the instructor is evaluated by their students. This finding has not been regularly discussed in the literature because the data on

individual student choice are hard to collect. This finding suggests that the class selection process is a *leading indicator* of the students' evaluation of the instructor. Inclusion of the *choice* variable should be considered a new contribution to the economics education literature.

Similarly, the native language variable is a new contribution to the student evaluation literature. Again, the sign of this variable may be anticipated, but the magnitude and statistical significance of this variable are noteworthy. This variable suggests that the students in this sample have a clear preference for those faculty members that share the students' native language. Since this variable *cannot* be controlled by the instructor, it should be considered when administrative decisions are made.

Finally, the gender match between the students and faculty increases the SET score in a way that corroborates other findings (e.g., Leuck, Endres and Caplen 1993). This finding has implications for administrative decisions, particularly for disciplines that normally have a high faculty-to-student gender match (e.g., nursing), or disciplines where gender match is changing (e.g., business).

1.3. Exam Findings

The achievement exam production functions estimated in this paper confirm the majority opinion in this area of research. Namely, student aptitude and effort are the most influential variables in a student's exam performance. Being older and being male also makes a positive contribution to performance. Furthermore, the teacher *does* seem to matter. Specific teaching attributes are shown to be influential in the SUR specification, whereas *general* teaching attributes are significant in the panel data specification. On the

other hand, institutional characteristics did not seem to influence student cognitive learning.

A contribution of these findings to the economics education literature comes not in the variables that are identified as predictors of cognitive learning, but rather for the way that the data were analyzed. Namely, previous studies of cognitive learning have generally taken class average exam scores and compared those to class average values for aptitude, effort, and so forth. This study has used individuals as the observation unit, rather than class sections. The results are very similar using either technique, and this demonstrates that the disaggregated learning equations perform similarly to the highly aggregated learning equations.

2. Caveats

This study has developed a faculty performance measurement system that requires several considerable changes from what has been called the "traditional" evaluation system. Among the changes required to implement this system are: (1) faculty agreement on a common outcomes test; (2) faculty and administrative agreement on the relative weights to give the SET and achievement exam scores; and (3) agreement on how often to estimate the prediction equation. There are others that could be added to this list. Because these are substantial changes from the current system, careful thought must be given to the implementation of this performance system. With that in mind, these results should be considered to be exploratory and preliminary findings. Another replication of this study would greatly enhance the ability to generalize these results.

Other warnings apply to any study of educational processes. First, as one studies education, it becomes apparent that no theory of learning applies to all situations. Although this study has addressed itself to teaching effectiveness, student learning cannot be directly controlled by the instructor, but rather the instructor can only facilitate learning. For that reason, there is an imperfect relationship between the effectiveness of the teacher and the learning outcomes of the students, and this is partially reflected in the low *explained* variation in learning that is reflected in the R^2 value of the learning equations. The reported R^2 values in this study are around 36 percent, which suggests that much of the explanation of these learning relationships remains unaccounted for. It must be noted, however, that these relatively low R^2 values are typical in cross-sectional research.

In the same vein, educational research findings are somewhat non-precise due to an incomplete understanding of student motivation. Although the teacher may teach well, student outcomes may not reflect good teaching owing to the possibility of a student setting a *target grade*, participating in extra-curricular activities, or other causes. Although several variables in this study have attempted to account for student effort, none of these variables have seemed capable of explaining these complex relationships.

Experimental design issues must be considered in the interpretation of this study. Because of the complex nature of education, additional explanatory variables in the learning equations may be a possibility for even richer findings. Some promising candidates for additional explanatory variables are: cognitive exam pretests, the weight of the achievement exam on a student's final grade, math aptitude, study time, a measure of family commitments, and possibly others.

The implications of repeated measures deserve discussion, too. Because this was a single study, *some* of the predictive relationships may cancel out in over several semesters. For example, the aptitude of the student was shown to be a strong predictor of achievement exam scores. In the long-run, however, the aptitude of the students in all classes will vary, and this variation may mean that *all instructors* get the same percentage of high-aptitude students when measured over a career. Other examples of variables that may cancel out over the long-run are ACT, GPA, BUS, and AGE, assuming that these variables were randomly distributed among instructors in the first place.[54] Nevertheless, while it may be true that some variables may cancel out in repeated measures, it is likely that others will *never* cancel out, such as the native language, choice, and expected grade variables. This suggests that controlling for these variables will always be appropriate.

Finally, grade inflation has been identified in this study, but not specifically controlled. Although a complete discussion of grade inflation is outside the scope of this study, some comments are appropriate. First, a prima facie case for the existence of grade inflation can be seen in the strong positive relationship between the grading pattern of certain instructors and their SET scores. Second, there is a consistent evidence of grade inflation in the economics education literature. Therefore, an attempt was made to identify an instructor's grading pattern in the teaching effectiveness measures developed in this paper, but there was no remedy proposed for grade inflation. Thus, an instructor could continue to be an "easy" grader, and thus get high student evaluations, but their

[54] The random effects model suggests that *there is* correlation between the right-hand side variables and the teachers, but the model gives no indication of *which* variables are correlated.

adjusted SET score would reflect their grading pattern. Perhaps by simply identifying grade inflation its practice would be reduced.

3. Policy Implications

Several possible uses of the composite index come to mind. First, as data base management improves over time, raw SET and exam scores could easily be adjusted using a prediction equation. This adjustment could be used on either a departmental or university-wide basis with appropriate control variables. These adjustments may help to silence the complaint that a certain professor got a "bad class," and his or her evaluations or exam results suffered as a consequence.

The faculty, often times suspicious of being evaluated by students, may also embrace a new method of evaluation because the proposed method gives a wider interpretation of classroom output. Thus, a faculty member who did not choose to "be popular" with students could be protected in this system by being identified for the outstanding cognitive learning of their students. Conversely, since it has been shown that the production of these two types of learning may require different instructional inputs, those instructors that chose to produce affective outputs would also be proportionately rewarded. Administrators would have many possible weighting schemes to value each skill.

Similarly, administrators must take seriously any system that discourages grade inflation in the pursuit of higher student evaluations. Since it is clear that easier graders get higher student evaluation ratings, ceteris paribus, any system to reduce this tendency is valued. The proposed system identifies and treats this grading bias.

In the final analysis, the use of student evaluations is an attempt to reduce the asymmetric information problem faced by administrators, but this practice has the unintended side effect of introducing a moral hazard problem. The broader standard offered by the composite index should serve to reduce the moral hazard problem that SETs introduce.

4. Suggestions for Future Research

Research in this area of economics education will continue because many interesting and nettlesome problems exist. Undoubtedly, research on better preparation of the faculty, the effect of new classroom technologies, and student motivation are areas rich in potential.

Some non-researched areas may emerge as well. One that comes to mind for economists is a study of the market-like attributes in the section-selection process of the student. That is, class sections are *not* typically rationed using the price system, as is common for other goods and services.[55] Instead, non-price rationing occurs as students select the *bundle* of teacher and institutional attributes in many of the same ways as a consumer selects a bundle of housing attributes. Students value teaching attributes such as the instructor's humor, grading policies, native language, intellectual rigor, among others. Also valued are institutional features such as time-of-day the class is offered, size of the class, and possibly others. As mentioned in this study, little research effort has been devoted to the student selection process. Instead, research has approached this problem

[55] Some colleges and universities have created tuition schemes based on student demand, but this is not the norm.

using the assumption that students are randomly assigned to their sections, and students react to the situation accordingly. A reversal of this paradigm might be productive. Namely, what if it were assumed that classes were selected in much the same way as a box of cereal is taken off a grocer's shelf? If the selection process reflected the non-price rationing of classroom space, and if information about the instructor is widely available, then the preference for a particular instructor would *theoretically* be reflecting the effectiveness of the teacher many months before the SETs were administered. This line of reasoning, taken to its logical conclusion, might make the measurement of learning outcomes unnecessary as an indicator of teaching effectiveness.

APPENDIX

Martin Chair of Insurance

P.O. Box 165
Middle Tennessee State University
Murfreesboro, Tennessee 37132
(615) 898-2673

APPENDIX 1

TO: Mark Wilson

FROM: Kenneth W. Hollman
Chair of Insurance

RE: Proposal entitled "An Examination of the Assessment Properties of the Faculty Evaluation Process": Protocol #97067

DATE: November 19, 1996

Today I signed the Institutional Review Board approval form for the above referenced proposal. The approval is good for one year. If you see that you cannot complete the study within a year, you should ask for a renewal prior to the end of the year. You may seek approval for only two additional years.

Please contact me at 898-2673 or Ms. Myra Norman at 898-5005 if you have a question. Thank you for your patience.

C: Dr. J. Zietz
Dr. Myra Norman

APPENDIX 2

SIGN NAME:_____ **DATE:**_____

I authorize this information to be used in an MTSU study of the teacher evaluation process. I understand that I will not personally be identified in any way.

	CIRCLE CHOICE
1) How do you rate the overall performance of your instructor in this course?	a) outstanding b) above average c) average d) below average e) poor
2) Do you have a job this semester?	YES or NO
3) Average hours worked per week:	_____hours
4) Is your major in the college of Business?	YES or NO
5) Is this a required or elective class for you?	REQUIRED or ELECTIVE
6) When you registered, was this your first choice of class time?	YES or NO
7) Your marital status:	MARRIED or UNMARRIED
8) Your age:	_____years
9) Your current GPA:	_____gpa
10) Your expected grade in this class:	A B C D F
11) Your gender:	MALE or FEMALE
12) Your class rank:	FR SOPH JUN SR
13) Did you take high school economics?	YES or NO
14) Before this class, how many college-level economics courses had you taken?	_____courses

THANK YOU

APPENDIX 3

PRINCIPLES OF MACROECONOMICS **NAME:**_____

COMPREHENSIVE EXAM **December, 1996**

INSTRUCTIONS: Select the best answer to all of the following questions.

1. Inflation can be defined as:
   a) a sustained increase in production of goods and services
   b) a sustained increase in the average price level
   c) a one-time increase in the income tax rate
   d) any increase in the price of a key good or service

2. Gross domestic product is:
   a) the largest industry in any developed economy
   b) the volume of goods and services involved in international trade
   c) non-market production of goods and services
   d) the money value of all goods and services produced in a given year

3. The fundamental issue of economics is to:
   a) give every citizen good housing and medical care
   b) redistribute income and wealth to eliminate poverty
   c) reduce unemployment and thus welfare payments
   d) learn to cope with the scarcity of virtually all resources

4. Which of the following pairs of outputs would most likely have a straight-line Production Possibilities Frontier?
   a) oil and natural gas
   b) football playing and basketball playing
   c) blue cars and red cars
   d) cars and airplanes

5. Questions of what to produce, how to produce, and who will get the output are faced by:
   a) market economies
   b) "emerging" economies
   c) centrally planned economies
   d) all economies

6. Which of the following is correct?
   a) MPC - MPS = 1.0
   b) 1.0 - MPC = MPS
   c) 1.0 - MPC = 1.0 - MPS
   d) MPC + MPS = APC - APS

7. An economy that wants to experience economic growth should:
   a) increase current production of consumer goods, reduce current production of capital goods
   b) increase current production of capital goods, reduce current production of consumer goods
   c) increase its current production of both consumption and capital goods
   d) decrease its current production of both consumption and capital goods

8. The most significant real cost of inflation is that it:
   a) erodes the purchasing power of wages
   b) leads to unfairly high prices
   c) steals from the rich to aid the poor, distorting incentives
   d) redistributes income in an arbitrary way

9. A fairly common cause of recession and unemployment is:
   a) insufficient aggregate demand
   b) excess aggregate demand
   c) insufficient aggregate supply
   d) excess aggregate supply

10. In terms of the aggregate supply and demand diagram, if AD shifts to the right:
   a) both prices and nominal GDP will rise
   b) both prices and nominal GDP will fall
   c) prices and nominal GDP can either rise or fall; more information needed
   d) nominal GDP will rise and prices may either rise or fall; more information needed

11. Persons who do not have jobs and who do not look for work are considered:
   a) unemployed
   b) out of the labor force
   c) "underemployed" in the employment statistics
   d) part of the "underground" economy engaged in illegal activities

12. Why do lenders lose during times of unexpected inflation?
   a) they are unable to keep up with prices
   b) everyone loses when inflation occurs
   c) they have fixed incomes
   d) they receive money that has less purchasing power than the money they lent out

13. What does it mean if the price index is 122.3 this year and 100 in the base year?
   a) what can be bought for $100 today cost 122.30 in the base year
   b) prices have increased on average by 122.3% since the base year
   c) output is 22.3% higher than in the base year
   d) prices have increased on average by 22.3% since the base year

14. An appreciation of the exchange rate will most likely:
    a) worsen the trade deficit
    b) reduce the trade deficit
    c) raise exports and imports
    d) lower exports and imports

15. A certain measure of income is defined as the sum of incomes of all individuals in the economy after taxes have been deducted and transfer payments added in. This is:
    a) disposable income
    b) national income
    c) capital income
    d) nominalized real income

16. With a constant government budget deficit, an increase in investment requires:
    a) private and foreign savings to fall
    b) a rise in private savings and a fall in the trade deficit
    c) a rise in private and/or foreign savings
    d) none of the above

17. If one compares the variability of investment spending to consumption spending:
    a) investment spending is more variable
    b) consumer spending is more variable
    c) they are about equally variable
    d) it is not possible to generalize about their variability

18. The rate of interest will affect investment spending because:
    a) much investment is financed by borrowing
    b) interest rates affect bond prices
    c) the stock market falls or rises directly with interest rates
    d) the interest rate is the only determinant of saving

19. In a fully-employed economy the income multiplier associated with increased government purchases is:
    a) greater than in an economy facing unemployment
    b) smaller than in an economy facing unemployment
    c) equal to the one for an economy facing unemployment
    d) not determined

20. According to the "balanced budget multiplier," equal increases in government spending and taxation will:
    a) raise the level of income
    b) lower the level of income
    c) not change the level of income
    d) lower interest rates as well as lower the income

21. A farmer earns $70,000, spends $40,000 on personal goods, and $30,000 on a new tractor. He has:
    a) consumed $70,000
    b) saved $0
    c) invested $30,000
    d) saved $40,000

22. An increase in U.S. interest rates is likely to trigger:
    a) a capital inflow and a dollar depreciation
    b) a capital outflow and a dollar depreciation
    c) a capital inflow and a dollar appreciation
    d) a capital outflow and a dollar appreciation

23. Specializing and trading on the basis of comparative advantage:
    a) decreases a country's production possibilities
    b) increases a country's consumption and production possibilities
    c) expands a country's consumption possibilities
    d) expands a country's production possibilities

24. Assume the income multiplier is 5 and equilibrium GDP is $500 billion below full-employment GDP. The amount of additional spending to bring the economy up to full-employment GDP is:
    a) 1,000 billion
    b) 500 billion
    c) 250 billion
    d) 100 billion

25. An increase in wages will cause the aggregate supply curve to:
    a) shift upward
    b) shift downward
    c) become steeper
    d) become flatter

26. To pay back its international debt, a developing country needs to increase its:
    a) consumption
    b) domestic money supply
    c) rate of population growth
    d) export potential

27. Currently, money in the U.S. is backed by:
    a) gold stored (primarily) at Ft. Knox, Kentucky
    b) gold and silver in government vaults
    c) a willingness of people to accept it for transactions
    d) Federal Reserve Notes, and the ability of the Fed to redeem those in silver

28. Banks raise the money supply when they:
    a) put more coins into circulation
    b) print currency
    c) increase their assets by charging interest
    d) convert their excess reserves into loans

29. What will happen to the money supply if the Fed lowers the discount rate?
    a) money supply will increase
    b) money supply will decrease
    c) there will be no change unless the money multiplier changes
    d) the money supply will remain the same but the velocity of circulation will increase

30. GDP per capita is barely growing in many developing countries because:
    a) their output of goods and services is steadily decreasing
    b) they possess no comparative advantage in production
    c) their populations are increasing about as fast as real GDP
    d) there are no technology transfers from industrial countries

APPENDIX 4

PRINCIPLES OF MICROECONOMICS **NAME:**_____
COMPREHENSIVE EXAM **December, 1996**

INSTRUCTIONS: Select the best answer to all of the following questions.

1. In 1971, a bank teller could process 265 checks per hour. By 1988, computer scanners had increased that figure to 825 checks per hour. Economists describe this type of change as:

   a) economic growth
   b) increase in labor productivity
   c) comparative advantage
   d) gains from voluntary trade

2. Which of the following is an example of efficient specialization and voluntary trade?

   a) A college professor hires someone to till a garden in the spring
   b) A college professor works on the engines of a car for a neighbor who is a mechanic
   c) A lawyer decides to baby-sit his young child and agrees to do so for others for cash
   d) a physician agrees to help a neighbor do her taxes in exchange for bookkeeping services

3. The fundamental issue of economics is to:

   a) reduce shortages
   b) redistribute income to eliminate poverty
   c) reduce unemployment and thus welfare payments
   d) learn to cope with scarcity of virtually all resources

4. Which of the following would be most likely to cause an outward shift of the demand curve for electricity?

   a) a decrease in the price of electricity
   b) an increase in the price of air conditioners
   c) an increase in the price of heating oil
   d) a decrease in the price of natural gas

5. Normally, an increase in the supply of a good will cause:

   a) a shift in consumer preferences in favor of that good
   b) consumers to use more of that good and less of others
   c) a shift in consumer preferences away from that good
   d) consumers to use less of that good and more of others

6. The slope of the demand curve is almost always:

a) positive, because when people buy more of a good, the cost of producing it will rise
b) positive, because the more money a person has, the more of a particular good will be bought
c) negative, because with all else equal, the same people will buy more of a good at a lower price
d) negative, because when people buy more of a good the price of producing it will fall

7. Contrasted with uncontrolled prices, prices in a "black market" are usually:

a) lower, since it is hard for the sellers to locate buyers
b) lower, since it is hard for the buyers to locate the sellers
c) higher, since black marketers expect compensation for the risk of being caught
d) higher, since most people enjoy the goods more if they are illegal

8. Meat from Porky the Pig can be used to produce bacon or sausage, but not both. If the price of bacon rises for some reason, then, all else equal:

a) the price of sausage will rise
b) the price of sausage will fall
c) the resources used raising Porky will become more expensive
d) the resources used raising Porky will become less expensive

9. Profits are maximized only where marginal revenue equals:

a) total cost
b) marginal cost
c) average revenue
d) total revenue

10. If over some range of production average cost is falling, the firm is experiencing:

a) increasing returns to scale
b) decreasing returns to scale
c) constant returns to scale
d) increasing costs per unit of output

11. A production function:

a) gives the maximum output that can be obtained from each combination of inputs
b) is derived from the marginal physical product curve
c) is derived from the cost curve
d) defines the profit maximizing input combination for a firm

12. A price cut will reduce the revenue a firm receives if the demand for its products is:

a) elastic
b) inelastic
c) unit elastic
d) pre-elastic

13. A five cent tax per gallon of gasoline would raise the price paid by consumers by:

a) more than five cents
b) five cents
c) less than five cents
d) five cents times the price elasticity of demand

14. The long-run price elasticity of demand is generally larger than the short-run elasticity because:

a) it takes people time to realize that prices have risen
b) peoples' income changes over time
c) it takes time to find substitutes
d) people need time to discover flaws in a product

15. The demand for labor is a "derived" demand. Employers hire workers until the:

a) last worker adds nothing to total output
b) average product is zero
c) wage rate equals the marginal revenue product of labor
d) wage rate equals the average product of labor

16. An "inferior" good is one:

a) whose price falls as incomes increase
b) whose quantity demanded falls when the purchaser's income rises
c) whose demand increases only with increases in income
d) whose market price is unaffected by income changes

17. Market demand curves are found by:

a) adding income levels for given price levels
b) horizontally summing individual demand curves
c) summing individual demand curves in a parallel fashion
c) adding the slopes of individual demand curves

18. Firms will continue to enter a competitive industry until:

a) the supply curve is vertical
b) all resources are fully employed
c) accounting profits are zero
d) any excess returns have been competed away

19. One of the following is NOT a characteristic of perfect competition. Which one?

a) firms advertise to increase their market share
b) profits are low in the long run
c) firms pay no attention to the behavior of their rivals
d) consumers pay little attention to brand names

20. Country 1 can produce 30 A or 40 B per resource input. Country 2 can produce 20 A or 30 B. Hence:

a) Country 1 has a comparative advantage in A and Country 2 in B
b) Country 1 has a comparative advantage in B and Country 2 in A
c) Country 1 has a comparative advantage in both A and B
d) Country 2 has a comparative advantage in both A and B

21. Prices serve the public interest by:

a) raising the rate of return earned by resource owners
b) allocating scarce resources
c) keeping poor people from buying more than they can afford
d) forcing the government to participate in the market

22. A monopolist maximizes profits by producing at which of the following?

a) MC = P
b) AC = P
c) AC = AR
d) MC = MR

23. When social marginal costs exceed private marginal costs, a company will:

a) overproduce
b) underproduce
c) reduce output
d) sell at a loss

24. The lack of property rights and/or the inability to enforce them is:

a) responsible for shortages
b) the fundamental reason for moral hazard
c) an implication of adverse selection
d) largely responsible for the decline of the American buffalo population

25. Monopolistic competition is common in:

a) retail selling
b) farming
c) electric power generation
d) telecommunications

26. The income effect of a higher wage:

a) leads to more work effort
b) leads to less work effort
c) is always less than the substitution effect
d) works in the same direction as the substitution effect

27. Tariffs on imports:

a) protect jobs in export-competing industries
b) increase foreign aggregate demand
c) benefit domestic industries that compete with imports
d) harm only a minority of the population

28. The demand curve for labor slopes downward because:

a) the substitution effect is greater than the income effect
b) capital has been substituted for labor in most industries
c) few workers will work at low wages
d) of the diminishing marginal product of labor

29. Which of the following distinguishes a monopolist from a monopolistically competitive industry?

a) entry barriers
b) a downward sloping demand curve for the firm
c) the size of the firm
d) the profit maximization rule is different for each

30. A progressive tax is one for which the % of each added dollar of income paid in taxes:

a) increases as income increases
b) remains the same as income increases
c) decreases as income increases
d) is zero after maximum income is reached

BIBLIOGRAPHY

BIBLIOGRAPHY

Aigner, Dennis J., and Frederick D. Thum. "On Student Evaluations of Teaching Ability," *Journal of Economic Education*, Fall 1986, pp. 243-265.

Akerlof, George. "The Market for Lemons," *Quarterly Journal of Economics*, August 1970, pp. 488-500.

Alchian, Armen A., James M. Buchanan, Harold Demsetz, Axel Leijonhufved, John R. Lott, Williams F. Sharpe, and Robert H. Topel. "In Celebration of Armen Alchian's 80th Birthday: Living and Breathing Economics," *Economic Inquiry*, July 1996, pp. 412-426.

Battalio, R.C., J. R. Hulett, and J.H. Kagel. "Comment on J.J. Siegfried's 'The Publishing of Economic Papers and Its Impact on Graduate Faculty Ratings, 1960-1969'," *Journal of Economic Literature*, March 1973, pp. 68-70.

Becker, William E. "The University Professor As a Utility Maximizer and Producer of Learning, Research and Income," *The Journal of Human Resources*, Winter 1975, pp. 107-115.

_____. "Professorial Behavior Given a Stochastic Reward Structure," *American Economic Review*, December 1979, pp. 1010-1017.

_____. "The Educational Process and Student Achievement Given Uncertainty in Measurement," *American Economic Review*, March 1982, pp. 229-236.

_____. "Economic Estimation Research: Part III, Statistical Estimation Methods," *Journal of Economic Education*, Summer 1983, pp. 4-15.

_____. "Teaching Economics to Undergraduates," *Journal of Economic Literature*, September 1997, pp. 1347-1373.

Blackwell, J. Lloyd. "A Statistical Interpretation of Student Evaluation Feedback: A Comment," *Journal of Economic Education*, Summer 1983, pp. 28-31.

Borg, Mary O., Paul M. Mason, and Stephen L. Shapiro. "The Case of Effort Variables in Student Performance," *Journal of Economic Education*, Summer 1989, pp. 308-313.

Bosshardt, William, and Michael Watts. "Instructor Effects in Economics in Elementary and Junior High Schools," *Journal of Economic Education*, Summer 1994, pp. 95-211.

Boyle, R. P. "Path Analysis and Ordinal Data," *American Journal of Sociology*, January 1970, pp. 461-480.

Brasfield, David W., Dannie E. Harrison, and James P. McCoy. "The Impact of High School Economics on the College Principles of Economics Course," *Journal of Economic Education*, Spring 1993, pp. 99-111.

Broder, Josef M., and William J. Taylor. "Teaching Evaluation in Agricultural Economics and Related Departments," *American Journal of Agricultural Economics*, February 1994, pp. 153-162.

Brown, Gregory A. *Course Length as a Determinant of Student Performance in the Principles of Macroeconomics Course*, D.A. dissertation, Middle Tennessee State University, 1996.

Cappoza, Dennis. "Student Evaluations, Grades, and Learning in Economics 912," *Western Economic Journal*, March 1973, p. 127.

Centra, John A. *Reflective Faculty Evaluation*, 1993, San Francisco: Jossey-Bass Publishers.

Chizmar, John F., and Thomas A. Zak. "Modeling Multiple Outputs in Educational Production Functions," *American Economic Review*, May 1983, pp. 18-22.

Chizmar, John F., and David E. Spencer. "Testing the Specification of Economic Learning Equations," *Journal of Economic Education*, Spring 1980, pp. 45-49.

Cochran, Howard Henry, Jr. *The Influence of Class Size on Student Achievement in Principles of College Economics: A Production Function Approach*, D.A. dissertation, Middle Tennessee State University, 1994.

Costin, Frank, William T. Greenough, and Robert J. Menges. "Student Ratings of College Teaching: Reliability, Validity, and Usefulness," *Review of Educational Research*, December 1971, pp. 511-535.

Davisson, William I., and Frank J. Bonello. *Computer-assisted instruction in economics education: A case study*, 1976, Notre Dame, IN: University of Notre Dame Press.

DeCanio, Stephen J. "Student Evaluations of Teaching--A Multinomial Logit Approach," *Journal of Economic Education*, Summer 1986, pp. 165-176.

DeMeza, David, and Michael Osborne. *Problems in Price Theory*, 1980, Chicago: University of Chicago Press.

Dennis, Louise I. "Student Evaluations: Are They an Appropriate Criterion for Promotion?", *Nursing Health Care*, February 1990, pp. 79-82.

Dilts, David A. "A Statistical Interpretation of Student Evaluation Feedback," *Journal of Economic Education*, Spring 1980, pp. 10-15.

_____, and Ali Fatemi. "Student Evaluation of Instructors: Investment or Moral Hazard?", *Journal of Financial Education*, Fall 1982, pp. 67-70.

Douglas, Stratford, and Joseph Sulock. "Estimating Educational Production Functions with Correction for Drops," *Journal of Economic Education*, Spring 1995, pp. 101-112.

Everett, Michael D. "Student Evaluations of Teaching and the Cognitive Level of Economic Courses," *Journal of Economic Education*, Spring 1977, pp. 100-103.

_____. "The Impact of Student Evaluations of Teaching on the Quality of Education," *Liberal Education*, Winter 1981, pp. 327-335.

Gibbons, Jean D., and Mary Fish. "Rankings of Economics Faculties and Representation on Editorial Boards of Top Journals," *Journal of Economic Education*, Fall 1991, pp. 361-372.

Goldstein, Robert Justin. "Some Thoughts about Standardized Teaching Evaluations," *Perspectives on Political Science*, Winter 1993, pp. 8-11.

Gramlich, Edward M., and Glen A. Greenlee. "Measuring Teacher Performance," *Journal of Economic Education*, Winter 1993, pp. 3-13.

Greene, William H. *Econometric Analysis: Second Edition*, 1993, New York: Macmillan Publishing Company.

Hall, Bronwyn H., and Clint Cummins. *Times Series Processor Version 4.3 User's Guide*, 1995, Palo Alto, CA: TSP International.

Hansen, W. Lee, and Allen C. Kelley. "Political Economy of Course Evaluations," *Journal of Economic Education*, Fall 1973, pp. 10-21.

Hanushek, Eric A. "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, September 1986, pp. 1141-1177.

Heckman, J.J. "Sample Selection Bias as a Specification Error," *Econometrica*, January 1979, pp. 153-161.

Highsmith, Robert J., and William J. Baumol. "Education in Economics: Evidence on Determinants of Effectiveness," *American Journal of Agricultural Economics*, December 1991, pp. 1380-1385.

Jacobs, Lucy Chester, and Clinton I. Chase. *Developing and Using Tests Effectively*, 1992, San Francisco: Jossey-Bass Publishers.

Judge, G.C., W.E. Griffiths, R. Carter Hill, H. Lutkepohl, and Tsoung-Chao Lee. *The Theory and Practice of Econometrics*, 1985, New York: John Wiley and Sons.

Kau, James B., and Paul H. Rubin. "Measurement Techniques, Grades, and Ratings of Instructors," *Journal of Economic Education*, Fall 1976, pp. 57-62.

Kelley, Allen C. "TIPS and Technical Change in Classroom Instruction," *American Economic Review*, May 1972, pp. 422-428.

_____. "Uses and Abuses of Course Evaluations as Measures of Educational Output," *Journal of Economic Education*, Fall 1972, pp. 13-18.

_____. "The Student as a Utility Maximizer," *Journal of Economic Education*, Spring 1975, pp. 82-92.

Kipps, Paul H. "The Use of Course Evaluation Scores to Influence Teaching and Research Activities," *Journal of Economic Education*, Spring 1975, pp. 93-98.

Labovitz, S. "The Assignment of Numbers to Rank Order Categories," *American Sociological Review*, June 1970, pp. 515-524.

Leventhal, Les, Philip C. Abrami, Raymond P. Perry, and Lawrence J. Breen. "Section Selection for the Validation and Use of Teacher Rating Forms," *Educational and Psychological Measurement*, Winter 1975, pp. 885-895.

Lichty, Richard W., David A. Vose, and Jerrold M. Peterson. "The Economic Effects of Grade Inflation on Instructor Evaluation: An Alternative Interpretation," *Journal of Economic Education*, Fall 1978, pp. 3-11.

Lima, Anthony K. "An Economic Model of Teaching Effectiveness," *American Economic Review*, December 1981, pp. 1056-1059.

Lopus, Jane S., and Nan L. Maxwell. "Beyond High School: Does the High School Economics Curriculum Make a Difference?", *The American Economist*, Spring 1994, pp. 62-69.

_____. "Teaching Tools: Should We Teach Microeconomic Principles Before Macroeconomic Principles?", *Economic Inquiry*, April 1995, pp. 336-350.

Lueck, Therese L., Kathleen L. Endres, and Richard E. Caplan. "The Interaction Effects of Gender and Teaching Evaluations," *Journalism Educator*, Autumn 1993, pp. 46-54.

MacDowell, Michael A., Peter R. Senn, and John C. Soper. "Does Sex Really Matter?", *Journal of Economic Education*, Fall 1977, pp. 28-33.

Machina, Kenton. "Evaluating Student Evaluations," *Academe*, May/July 1987, pp. 19-22.

Machlup, Fritz. "Poor Learning from Good Teachers," *Academe*, October 1979, pp. 376-380.

Manahan, Jerry. "An Educational Production Function for Principles of Economics," *Journal of Economic Education*, Spring 1983, pp. 11-16.

Marlin Jr., James W., and James F. Niss. "End-of-Course Evaluations as Indicators of Student Learning and Instructor Effectiveness," *Journal of Economic Education*, Spring 1980, pp. 16-27.

Mason, Paul M., Jeffrey W. Steagall, and Michael M. Fabritius. "Student Evaluations of Faculty: A New Procedure for Using Aggregate Measures of Performance," *Economics of Education Review*, 14:4, 1995, pp. 406-416.

Mason, Robert D. *Statistical Techniques in Business and Economics: Sixth Edition*, 1986, Homewood, IL: Richard D. Irwin, Inc.

Maxwell, Nan L., and Jane S. Lopus. "The Lake Wobegon Effect in Student Self-Reported Data," *American Economic Review*, May 1994, pp. 201-205.

McConnell, Campbell R., and Stanley L. Brue. *Contemporary Labor Economics, Fourth Edition*, 1995, New York: McGraw-Hill, Inc.

McCoy, James P., Don Chamberlain, and Rob Seay. "The Status and Perception of University Outcomes Assessment in Economics," *Journal of Economic Education*, Fall 1994, pp. 358-366.

McKeachie, Wilbert J. "Student Rating of Faculty: A Reprise," *Academe*, October 1979, pp. 384-397.

McKenzie, Richard B. "The Economics of Reducing Faculty Teaching Loads," *Journal of Political Economy*, June 1972, pp. 617-619.

Mirus, Rolf. "Some Implications of Student Evaluations of Teachers," *Journal of Economic Education*, Fall 1973, pp. 35-37.

Morgan, W. Douglas, and Jon David Vasche. "An Educational Production Function Approach to Teaching Effectiveness," *Journal of Economic Education*, Spring 1978, pp. 123-126.

Mulligan, James G. "A Classroom Production Function," *Economic Inquiry*, April 1984, pp. 218-226.

Needham, Douglas. "Student Effort, Learning and Course Evaluation," *Journal of Economic Education*, Fall 1978, pp. 35-43.

_____. "The Economics of Reducing Faculty Teaching Loads: Comment," *Journal of Political Economy*, February 1975, pp. 219-223.

Nelson, Jon P., and Kathleen Lynch. "Grade Inflation, Real Income, Simultaneity, and Teaching Evaluations," *Journal of Economic Education*, Winter 1984, pp. 21-37.

Nichols, Alan, and John C. Soper. "Economic Man in the Classroom," *Journal of Political Economy*, September/October 1972, pp. 1069-1073.

Park, Kang H., and Peter M. Kerr. "Determinants of Academic Performance: A Multinomial Logit Approach," *Journal of Economic Education*, Spring 1990, pp. 101-111.

Rodin, Miriam, and Burton Rodin. "Student Evaluations of Teachers," *Journal of Economic Education*, Fall 1973, pp. 5-9.

Rose, Louis A. "Adjustment of Student Ratings of Teachers for Extrinsic Influences," *Journal of Economic Education*, Spring 1975, pp. 129-132.

Sabot, Richard, and John Wakeman-Linn. "Grade Inflation and Course Choice," *Journal of Economic Perspectives*, Winter 1991, pp. 159-170.

Saunders, Phillip. "Learning Theory and Instructional Objectives," in *The Principles of Economics Course*, Phillip Saunders and William B. Walstad, 1990, New York: McGraw-Hill Publishing Company, pp. 62-83.

Schmidt, Robert M. "Who Maximizes What? A Study in Student Time Allocation," *American Economic Review*, May 1983, pp. 23-28.

Scott, Robert C. "A Comment on 'Grade Inflation: A Way Out,'" *Journal of Economic Education*, Summer 1988, p. 227.

Seiver, Daniel A. "Evaluations and Grades: A Simultaneous Framework," *Journal of Economic Education*, Summer 1982, pp. 32-38.

Shmanske, Stephen. "On the Measurement of Teacher Effectiveness," *Journal of Economic Education*, Fall 1988, pp. 307-314.

Siegfried, John J. "A Reply to the Comment of Professors Battalio, Hulett, and Kagel on 'The Publishing of Economic Papers and Its Impact on Graduate Faculty Ratings, 1960-1969,'" *Journal of Economic Literature*, March 1973, pp. 71-73.

_____, and Rendigs Fels. "Research on Teaching College Economics: A Survey," *Journal of Economic Literature*, September 1979, pp. 923-969.

Sowell, Thomas. *Inside American Education*, 1993, New York: Free Press.

Stratton, Richard W., Steven C. Myers, and Randall H. King. "Faculty Behavior, Grades, and Student Evaluations," *Journal of Economic Education*, Winter 1994, pp. 5-15.

Tuckman, Howard P. "Teacher Effectiveness and Student Performance," *Journal of Economic Education*, Fall 1975, pp. 34-39.

Valenzuela, Angela, and Sanford M. Dornbusch. "Familism and Social Capital in the Academic Achievement of Mexican Origin and Anglo Adolescents," *Social Science Quarterly*, March 1994, pp. 18-36.

Watts, Michael, and Gerald J. Lynch. "The Principles Courses Revisited," *American Economic Review*, May 1989, pp. 236-241.

Watts, Michael, and William Bosshardt. "How Instructors Make a Difference," *The Review of Economics and Statistics*, May 1991, pp. 336-340.

Weidenaar, Dennis, J., and Joe A. Dodson, Jr. "The Effectiveness of Economics Instruction in Two-Year Colleges," *Journal of Economic Education*, Fall 1972, pp. 5-12.

Wetzstein, Michael E., Joseph M. Broder, and Gene Wilson. "Bayesian Inference and Student Evaluations of Teachers and Courses," *Journal of Economic Education*, Winter 1984, pp. 40-45.

White, Lawrence J. "Efforts by Departments of Economics to Assess Teaching Effectiveness: Results of an Informal Survey," *Journal of Economic Education*, Winter 1995, pp. 81-85.

Williams, Reed G., and John E. Ware, Jr. "Validity of Student Ratings of Instruction Under Different Incentive Conditions: A Further Study of the Dr. Fox Effect," *Journal of Education Psychology*, 68:1, 1976, pp. 48-56.

Yunker, James A., and James W. Marlin, Jr. "Performance Evaluation of College and University Faculty," *Educational Administration Quarterly*, Winter 1984, pp. 9-37.

Zangenehzadeh, Hamid. "Grade Inflation: A Way Out," *Journal of Economic Education*, Winter 1988, pp. 217-226.

Zietz, Joachim, and Howard H. Cochran, Jr. "A Note on Assessing Teaching Effectiveness," Unpublished Manuscript, Middle Tennessee State University Department of Economics and Finance, June 1996.