Money Only Matters if You Want It To? Exposing the Normative Implications of Empirical Research

KEVIN B. SMITH AND J. SCOTT GRANBERG-RADEMACKER, UNIVERSITY OF NEBRASKA AT LINCOLN

Political scientists have widely employed production function models as tools for theory confirmation and policy prescription. Although an important part of a growing literature within the discipline, production function research frequently produces contradictory results, an inconsistency that raises questions about the role of normative preferences in quantitative analysis. In this article we seek to explain the variation in the empirical results of production functions research in education. While recognizing normative values may have some influence in research, we argue this can be minimized if the conceptual and methodological weaknesses inherent in applying production function logic to the public sector are recognized and addressed.

Academic educational policy analysis has been widely used to justify prescriptive reforms of the public school system and as a laboratory to seek confirmation for broad theoretical frameworks such as public choice and representative bureaucracy (e.g., Chubb and Moe 1988, 1990; Henig 1994; Smith and Meier 1995; Meier, Wrinkle, and Polinard 1999; Maranto, Milliman, and Stevens 2000; Witte 2000; Weiher 2000; Schneider, Teske, and Marschall 2000). A significant portion of this research is methodologically moored to education production function analysis, a modeling approach that seeks to isolate the determinants of school outputs using correlation or regression techniques (Fortune 1993). Policy analyses using this approach, however, are notorious for producing inconsistent and contradictory results on what does (or does not) determine school performance (for surveys see Burtless 1996; Verstegen and King 1998). As a number of theoretical issues important to political science are now staked on such analyses, the discipline has an important interest in this question: What explains the variable and incongruent patterns reported in empirical studies of school performance?

In this article we argue that a good deal of the inconsistency is attributable to the conceptual problems created by exporting the analytic framework of production functions from economics and applying it to public policy and public agencies. We argue these problems can be minimized through a set of "second best solutions" that can be broadly applied to policy studies with little need for any methodological re-tooling by researchers.

The Logic and Practice Production Functions Analysis

The production function framework that logically structures a good deal of empirical policy analysis in political science originated in economics, and was designed for productivity and efficiency analysis of private firms. The best known formal treatment of the model is the Cobb-Douglas production function:

$$X = f(L,K) = L^a K^b$$

Here, a firm's output (X), is modeled as a function of L (labor) and K (capital). The relationship between these inputs and outputs is recovered through a and b, the parameters to be estimated. Given valid measures of output, labor and capital, the model allows easy calculation of the marginal impacts of the factors of production. The output measure is assumed to be directly related to the overarching goal of the firm. This goal is assumed to be profit maximization, and profit (P) is defined as: $P = cX - cL - cK$, where c represents costs. Given this, the production function model reduces questions of how to maximize a firm's central goal to a few straightforward calculations (specifically, the maximizing conditions for P are when $\partial P/\partial L = 0$, and $\partial P/\partial K = 0$). This provides a parsimonious, theoretically derived analytic framework to assess outputs and goal maximization for firms (for an accessible primer see Zellner et al. 1966).

The basic logic, if not the mathematical elegance, has long been exported to the rest of the social sciences. Typically economic theory is replaced with an alternate framework that intuitively "fits" the input-output logic of the Cobb-Douglas model. In political science this usually means substituting some unit of government for the firm, adopting some measure of agency or policy output such as expenditures or regulatory actions, and replacing labor and capital with a set of inputs causally linked to outputs by some throughput process undertaken at the unit of analysis. It is also common to impose linearity restrictions on the input-output relationship, thus making analysis more tractable using basic multi-variate regression techniques.[1]

[1] A good example in political science is the "mainstream model" of state-level policy outputs derived from Easton's (1965) systems theory of politics. Here policy outputs are modeled as function of inputs such as public opinion, state wealth, and partisan control of government. This

Political Research Quarterly, Vol. 56, No. 2 (June 2003): pp. 223-232

While this framework is flexible and widely employed by scholars interested in theoretical confirmation of agency- or policy-related hypotheses, the translation from firms to public agencies and policy entails a loss of conceptual clarity that imposes steep costs. Economists have long recognized a similar gap between econometric theory and econometric practice, with real-world data sets virtually dictating modeling approaches that drift from the parsimonious logic of basic production function theory. For example, firms pursue goals other than profit, and factors other than labor and capital (e.g., government regulation) obviously constrain production. How can these be accounted for? It is a relatively easy matter to employ dependent variables that measure goals other than profit, or to add any number of control variables to the right hand side of the equation. The problem with this common sense response to accounting for the complexities of the real world is that it not only loosens the theoretical reins on a study, but undermines the very basis of inference. A typical production function study regression analysis is reported as if it follows the well-known basic form:

$$Y = \alpha + \beta X + e \qquad (1)$$

Where Y is vector of observed output variables, X is a vector of variables theoretically presumed to determine Y, α and β are unknown parameters to be estimated in order to uncover the relationship between inputs (X) and outputs (Y), and e is a stochastic error term. The β's represent the critical information for purposes of theory confirmation or policy prescription, and they are presented as if they represent the entire test of the hypothesized relationship between the variables in X (inputs) and Y (outputs). As a number of scholars have pointed out, this is a misleading presentation of much empirical analysis using production function logic and regression methodologies (Gill 1999; Leamer 1978). A more accurate representation of the typical regression analysis is this:

$$Y = \alpha + \beta X + \gamma + e \qquad (2)$$

Where X represents a vector of variables that a researcher is reasonably convinced are causally related to Y, and Z represents a vector of variables whose relationship to Y is debatable and where the particular coefficients γ hold less theoretical or substantive interest than β. If an initial analysis produces a coefficient β that is "wrong" given a priori expectations, the model is ordinarily respecified by changing what is in included in vector Z in an attempt to get the model to exhibit the "right" fit with the data. What is left out of Z is, in effect, sent off to the error term and assumed away as a potential systematic cause of Y or of anything in X. Most reported regression analyses are a product of this generic specification decision-making process and are thus not unique, parsimonious models, but the *nth* version of an evolutionary process seeking a reasonably robust platform from which to survey and interpret coefficients on the variables of interest to the researcher (Frank 2000; Gill 1999; Leamer 1978: 64-67).

As Leamer (1978: 4) points out, such ad hoc specification searching not only creates a gap between research and theory, the mushy approach to variable inclusion and its treatment of the error term upends some of the assumptions necessary to support any meaningful inference from regression analysis. Coefficient direction and traditional tests of significance testing (t- or z-tests) are driven in part by the number and particular mix of variables used in a model. The inferences taken from any particular variable in a reported regression model thus may be less clear, or even contradicted, by alternate specifications that have not been reported. Leamer (1978: 122-24) argues the guilty secret of most quantitative empirical work is that it is a search for confirmation of subjectively determined prior expectations. In other words, the source of agreement or disagreement over any given model specification is as likely to be found in a disciplinary or individual preference for a particular outcome rather than in the axioms of theory or the technical details of methodology.

All of these problems increase in degree if not in kind when the target of empirical research is public policy or public agencies. Unlike firms, public agencies and policies lack any equivalent overarching goal equivalent to profit, and as goals are vague, multiple and complex, capturing these goals in output measures presents a formidable challenge (Wilson 1989; Meier 1997; Downs and Larkey 1986: 73; Wood and Waterman 1994). In addition to the increased difficulties of operationalizing outputs in the public sector, there also tend to be a wider spectrum of theoretically supportable inputs. In empirical studies of public policy or public agencies, what gets consigned to a stochastic error term or included as a predictor varies not just across variables but entire concepts. Compared to economics, the theoretical environment surrounding public agencies and policy reflects little in the way of a core set universal principles and is characterized by a diversity of analytic approaches compatible with a production function framework. The vagaries attributable to this conceptual pluralism are then wedded to the limitations of regression analysis. The lack of theoretical precision allows wide discretion in specifying causal linkages between concepts or variables, but once these choices are made they are implicitly postulated to meet the more demanding assumptions required to support the validity of inference from a regression model (Fortune 1993). Using the above notation as reference, one analyst's X is another analyst's Z, and while they may both be interested in Y, this is no guarantee that their output measures are tapping into the same concept. From a purely technical standpoint, that contradictory results are reported from such analyses should be no surprise.

IMPLICATIONS FOR EDUCATION PRODUCTION FUNCTION ANALYSIS

Education production functions have all the characteristics and the problems associated with the broader argument

basic framework dominated the empirical literature on state policy for decades (for a survey and examples see Treadway 1985).

made above. Typically, "education production function" means a garden-variety OLS regression seeking to isolate the correlates of a single indicator of school performance, and only the rare exception seeks the formal conditions of goal maximization (Fortune 1993). Such analyses rapidly part ways with the assumptions underpinning the Cobb-Douglas model. Schools have no clear overarching goal, and are expected to concurrently pursue a wide variety of often contradictory purposes. Prioritizing these to determine the overarching purpose(s) of public education is a normative philosophical, rather than an empirical or objective, exercise (see Tyack and Cuban 1995: 43).

Given the multiplicity of goals (instilling basic literacy, socializing future citizens, redressing social inequity etc.), it is not clear how—or what—output measures should operationalize them. Test scores, for example, might measure some element of a school's ability to instill basic skills, but higher SATs do not necessarily make a better citizen, nor are they necessarily reflective of equality of educational opportunity. There is even less clarity on the input side of education production function analyses. While researchers do face some more or less universal constraints on specifying inputs for education production function models (e.g., controls for socioeconomic characteristics), these tend to be more a convention grown from inductive trial and error rather dictated by any core theoretical principles. Beyond these basic constraints, however, there is a virtually endless list of potential variables that can (and probably have) been included in education policy analyses. Scholars attracted to education policy analysis come with sectarian theoretical interests with varying degrees of conceptual overlap. In political science these interests include applied policy analysis, empirical tests of theories of public choice and representative bureaucracy, and models of political socialization and participation (e.g., Miller 1981; Chubb and Moe 1990; Maranto, Milliman, and Stevens 2000; Meier and Stewart 1991; Nielsen and Wolf 2001). These substantive differences direct analysts toward differing sets of inputs—everything from a specific pedagogical practice to the racial composition of a school's faculty—and different outputs reflecting a range of school goals (academic excellence, equality of opportunity, etc.).

Given the diversity in theoretical perspectives, model specification, and operationalization of key concepts (not to mention data sources and unit of analysis), variation in what is (and is not) found to correlate with school performance is hardly surprising. And variation is certainly present in the political science literature. To mention but a few examples, there are contradictory conclusions on the impacts of institutional structure, forms of governance, size of bureaucracy, and faculty demographics on school performance (e.g., Chubb and Moe 1988, 1990; Smith and Meier 1995; Nielsen and Wolf 2001; Meier, Wrinkle, and Polinard 1999). Given the politically charged nature of proposals like vouchers or desegregation, it is easy to see how the inconsistent claims of empirical studies can take on a particular ideological tint (see Glenn 2001). These muddied conceptual waters seem to provide education production functions with a particularly aggressive buoyancy: they are largely incapable of conclusively sinking any reasonable hypothesis. The metaphor may be mixed, but it summarizes the basic problem.

This problem has two basic elements: (1) Education production functions tend to be vague about what they are trying to explain because much of the underlying theory is oriented toward vague concepts such as "school performance" rather than being goal specific. This means picking educational dimensions to test these theories is often a subjective exercise. Broad claims about theory confirmation or (especially) policy prescription on the basis of findings from a single educational output are suspect because such claims do not necessarily apply to any output beyond the particular measure that has been selected by the analyst or dictated by the constraints of a particular data set. They may not even be supported by alternate—and perhaps as theoretically defensible—models using the same dependent variable from the same data set. (2) The lack of theoretical precision exacerbates long known problems of operationalization, model selection and hypothesis testing. As Frank (2000: 149) points out, "The simple question, 'Yes, but have you controlled for xxx?' puts social scientists forever in a quandary as to inferring causality." The problem of controlling for confounding bias is succinctly illustrated by what is known as Simpson's Paradox: "Any statistical relationship between two variables may be reversed by including additional factors in the analysis" (Pearl 2000: 78). Pearl (2000: 136) argues that the assumptions needed to support causal inference from any coefficient β in a regression model are virtually ignored by the standard social science approach to policy analysis: "Econometric textbooks invariably devote most of their analysis to estimating structural parameters, but they rarely discuss the role of these parameters in policy evaluation." Academic policy analysts can ill afford such ignorance, especially in areas such as education where quantitative work is often used to justify prescriptive reforms. Model specification options in education production functions include a virtually endless set of possibilities for measuring and matching inputs to outputs, choices with implications for coefficient magnitude, direction and the outcomes of hypothesis tests. In making these operationalization choices we may miss our theoretical and applied targets even when they are clearly sighted.

These problems almost certainly provide enough gray area to allow for "normative creep"—the conscious or subconscious desire for a particular set of results—to influence the research process. As the choices in education production functions analysis are virtually endless, variation in results are hardly unexpected even from a purely technical point view. For practical purposes, the end result is the same from the perspective of a technical or a normative critique: a set of contradictory and mutually exclusive results that provide mixed messages for the purposes of theory confirmation and policy prescription. The real challenge is not explaining the contradictions generated by education production function studies, but in deciding what to do about them.

SECOND BEST SOLUTIONS

The most obvious solution to the problems we have sketched is a set of core theoretical principles. This would not completely insulate empirical research from the intrusion of normative values, but it would provide an explicit, universal set of theoretical targets, and clearer guidance for model specification. It would also set natural constraints on the claims made from empirical analyses aimed at one or just a handful of these goals. Unfortunately, such a theoretical core for educational policy analysis is unlikely. We have no complete theory of human behavior and how it is effected by educational environments, and the sheer diversity of theoretical interests attracted to education production function analysis is simply too broad to pretend one exists.

Lacking universal theoretical principles, the next best solution would be a methodology that would account for normative influences creeping unbidden into an analysis. Good methodological practice enforced through the peer review process reduces these problems, and there have been attempts to bring the normative elements of empirical research more explicitly into our methodologies (see Meier and Gill 2000). Nonetheless, expecting methodology to solve the underlying problems strikes us as unrealistic. Quantitative methods are tools for theory confirmation, and to attempt to substitute statistical approaches for the failings of theory is to argue for using a nail to drive a hammer. Lacking an applicable universal theory or methodology that automatically flags normative creep we argue for adopting two rules of thumb that might be termed "second best" solutions:

First, Account for the Multi-Dimensionally of Agency Outputs

Optimally, this should be done theoretically by conceptually clarifying an output in terms of an explicitly defined and justified agency goal. Given the risks of theoretical drift in operationalization, we would also suggest doing this empirically by using more than one output measure. There are several advantages to doing the latter: it explicitly recognizes a range of theoretical targets, variation in results may highlight what is important to different school missions, and thus clarify the tradeoffs necessary to achieve these goals. This would also provide a more realistic picture of how theoretical concepts of interest—governance, institutional structure, the racial composition of student bodies, specific policies and the like—fit into the complex environment of education.

We believe such an approach is especially important in education policy because schools experience economies of scope, i.e., they use their inputs to produce more than one good or goal and to produce more of one output they must reduce another (Wenger 2000). For example, beyond some point schools can only increase test scores by reducing graduation rates because pupils who perform badly on achievement tests are at higher risk of dropping out (Belfield and Levin 2002; Wenger 2000; Lankford and Wyckoff 1992). As Wenger (2000) points out, relatively few scholars have explicitly taken into account the economies of scope argument in educational policy analysis even though it offers a ready explanation for some of the contradictory findings in the literature. Rather than account for the multi-dimensionality of school outputs and explore the implications for theory confirmation and policy prescription scholars have been over-reliant on test scores as a generic measure of "school performance" or "quality education." For example, of the 377 production function studies included in Hanushek's (1997) survey, roughly three-quarters (282) used test scores as the dependent variable. Policy inference or hypothesis falsification taken from test score studies is not necessarily applicable to the wide range of alternate output measures with legitimate claim to be included under any umbrella term as broad as school performance. Indeed, the economies of scope argument suggests that policies boosting test scores may have negative impacts on other school goals.[2]

Second, Incorporate Measures of Coefficient Variability in an Analysis

Empirical analysis will remain vulnerable to charges of normative creep as long as it insists on the conceit that the reported model specification is the only reasonable option and that its results represent *in toto* the empirical test of the hypothesis under consideration. Coefficient direction, magnitude and statistical significance are dependent on specification, and the inconsistency of the education production functions literature provides a stark example of the potential effects of specification issues on causal inference. Even analyses securely anchored to theory typically involve an exploratory variable selection process that is generally unacknowledged (Gill 1999).

The problem of making causal inference from statistical analysis, of course, is not unique to education production functions, but the politically charged atmosphere surrounding them makes addressing this problem particularly acute. There are numerous potential responses to the problem. Some argue for abandoning the $p < .05$ alpha criterion, arguing it is an arbitrary threshold and that its underlying logic widely misunderstood. Proffered replacements include sophisticated sensitivity analyses, calculation of impact thresholds, or jettisoning traditional approaches entirely in favor of Bayesian techniques (for surveys see Frank 2000; Gill 1999). These make sense on their merits, but are often computationally intensive or require a methodological retooling that many are reluctant to undertake. Thus, we suggest reporting the variation of a coefficient across a range of theoretically defensible models using Leamer bounds.[3]

---

[2] It is not particularly clear what test scores do measure. They have been widely criticized as poor measures of educational aptitude, educational achievement, and human capital, and charged with inherent racial or cultural bias (e.g., Armstine 1995, Owen 1985, Rothstein 1997, Jencks 1998).

[3] Reporting variation across all potential permutations of a model is unrealistic and probably misleading. In a typical educational data set,

Suggested by Leamer (1983) and Leamer and Leonard (1983), Leamer bounds are a simple form of sensitivity analysis designed to expose the fragility of regression coefficients. Leamer bounds are simply the maximum and minimum values of a variable coefficient taken from all versions of a regression model run by an analyst. These require no new skills (just a little careful accounting) for those already comfortable with OLS regression. They are also easy to understand, the range providing an intuitive indication of a coefficient's stability under alternate specification scenarios. In effect, their purpose is to reveal the impact on β of the various combinations that can constitute Z in equation 2. When the minimum and maximum values of β in alternate specifications are on opposite sides of zero, the coefficient can be considered too fragile to be trustworthy regardless of its performance in the final model selected by an analyst (Gill 1999; Leamer 1983; Leamer and Leonard 1983). Including such measures achieves two goals. First, it reveals how dependent a given study's conclusions are on the particular set of specification choices made by the analyst. Second, in doing this it makes it harder for an analyst to "hide" a model that has been selectively specified to fit normative preconceptions.

Although neither of these second best solutions are adopted in the typical education production function analysis, combined we believe they offer the following advantages: (1) They channel an analyst toward greater conceptual clarity as a means to defend choice of output measure. (2) By using multiple output measures, i.e., different dependent variables, the analyst engages in the process of replication, in and of itself an important support to the validity of inferences taken from scientific inquiry (King 1995). (3) It places constraints on the potential normative drift of theory by making it harder to generalize results (in other words, using multiple dependent variables makes it more likely an analyst will have to face the explanatory limitations of her theory rather than making broad prescriptive claims from a test based on, say, a single test score output). (4) It provides some accountability for normative creep. If model selection is being driven by a desire for a particular outcome, high levels of coefficient variability are likely to appear unless this desire is also strongly reflected in the actual data. (5) Perhaps most importantly, by accounting for the multi-dimensional nature of school outputs, patterns across these dependent variables might provide information and insights that would be missed in a traditional study. Overall, the net effect will be to improve the validity of educational production functions as vehicles for theory confirmation, make them more accountable for normative creep (even if this cannot be entirely eliminated), and highlight systematic patterns between inputs and outputs that would otherwise be missed.

however, there will typically be several defensible options on how to operationalize, say, relative wealth, or financial resources (per pupil expenditures, revenue per student, taxable property per student, state support per student, etc.). What happens to the other coefficients as a result of these different choices? We are suggesting this is the sort of information that can and should be included as standard practice.

TABLE 1
DETERMINANTS OF TAAS PASS RATES IN TEXAS SCHOOL DISTRICTS

	Coefficient (Standard Error)
Core Variables	
Percent White	.049 (.016)
Percent Economically Disadvantaged	−.146 (.02)
Percent Gifted	.207 (.06)
Average Class Size (Student/Teacher Ratio)	−.42 (.11)
Teacher Experience	.30 (.08)
District Financial Independence (Percent Revenues From State)	−.03 (.01)
Test Variables	
Bureaucracy	
Percent Central Administration	−.665 (.19)
Percent School Administration	−.12 (.22)
Teacher/Professional Staff Ratio	.02 (.02)
Minority Representation	
Percent Black Teachers	−.25 (.03)
Percent Hispanic Teachers	−.01 (.01)
Competition	
Herfindahl Index	−.225 (1.16)
Resources	
Instructional Expenditures Per Pupil (× 1000)	.203 (.001)
N	935
R-Square	.41

Unstandardized Coefficient (Standard Error) reported

SECOND BEST PRODUCTION FUNCTIONS: AN EXAMPLE

Table 1 represents the results of an education production function typical of those published in the political science literature. The unit of analysis is Texas school districts in the 1998-99 school year, and the dependent variable is the percentage of students passing the Texas Assessment of Academic Skills (TAAS). The latter is a standardized test mandated for students in several grades and is used to assess

knowledge of the statewide curriculum. This dependent variable and similar data were employed in the recent exchange between Meier et al. (1999, 2001) and Nielsen and Wolf (2000), which centered on the implications of specification and related technical issues for Meier et al.'s (1999) claim that representative bureaucracy (more ethnically diverse teaching ranks) is positively correlated with minority and non-minority student performance. Other studies have employed comparable data with the TAAS measure as the dependent variable to examine the impact of bureaucracy and competition on school performance (e.g., Bohte 2001; Weiher 2000; Wrinkle et al. 1999). The "core variables" listed in Table 1 are those commonly employed as controls in such studies, and the "test variables" are those employed to test the hypotheses of interest in these and similar studies. For expository purposes we are treating expenditures as a test variable (in our terminology it is treated as a core variable in most other studies cited) because of its prominence in the education production functions literature. All these data are made publically available by the Texas Education agency and can be downloaded at http://www.tea.state.tx.us/perfreport/snapshot/.[4]

The findings for our core variables are relatively noncontroversial, and are repeated in other studies (e.g., Bohte 2001; Meier et al. 1999, 2001; Nielsen and Wolf 2000). Districts that have higher concentrations of Anglos, fewer students who are economically disadvantaged, greater proportions of students in gifted classes, smaller average class sizes, more experienced teachers, and are more financially independent (i.e., less dependent on state revenue), tend to have more students pass TAAS exams.

The findings for our test variables—i.e., those with implications for conflicting policy preferences—invite dissent. We used three variables to assess the impact of bureaucracy on school outputs: percent of district employees in central administration, percent in campus-level administration, and a "tooth-to-tail" ratio of the number of teachers per administrator (none of these were particularly highly correlated with each other and did not cause multi-collinearity problems—the highest VIFs in any of our models was about 4.0). The central administration variable is negative, supporting the argument that higher levels of bureaucracy constrain school outputs (Bohte 2001; Chubb and Moe 1990). The other two variables are insignificant, null findings that lend support to arguments that bureaucracy is more a scapegoat than a central cause of problems in education (Smith and Meier 1995). We used two variables to assess the impact of ethnic diversity of district faculty on outputs, the percent of teachers that were black and the percent of teachers that were Hispanic. Both had negative coefficients, the percent black being significant.[5] To test the impact of competition we employed a Herfindahl Index, a ratio that measures a district's share of the total enrollment in the county in which it is located (see Hoxby 2000: 1215 for a discussion of this measure's advantage in education studies).[6] The coefficient is insignificant and negative, a finding that echoes some other studies, but contradicts the primary policy inference generated by the public choice literature on education (Chubb and Moe 1990). Our finding for resources repeats the most commonly reported finding in the education production function literature—expenditures have a positive coefficient, but are statistically insignificant (Hanushek 1997; Verstegen and King 1998).

As is the case with all production function studies, analysts with a variety of theoretical and policy inferences are likely to question the validity of these results. Many would point out with perfect accuracy that our findings might change under alternate specification scenarios, scenarios that are just as theoretically defensible as that presented in Table 1. We do not contest such claims. As a demonstration of the utility of second best solutions, we deliberately seek to explore them.

This is done first by examining how the model performs with alternate output measures. We reasoned that there were at least three explicit educational goals that could be tested given the data available: knowledge transference, equality of educational opportunity, and socialization into good citizenship. All three have been championed as crucial educational goals (Rebell 1998; Tyack and Cuban 1995; Becker 1993; Coleman et al. 1966). Though cognizant of dissenting views (see Jencks 1998; Rothstein 1997), we judged test scores to be a reasonable measure of knowledge transference. TAAS pass rates are one way to do this, but the data set also offers another—the percent of students meeting "college criterion" on SAT or ACT tests (scores of 1110 on the SAT or 24 on the ACT).

Test scores can also be used to operationalize equality of educational opportunity because, regardless of their merits as indexes of achievement, exams like the TAAS, SAT and ACT are undoubtedly used to distribute social and economic

[4] We found a number of errors in the data. Some were easily correctable (e.g., a reported student/teacher ratio of 126 that turned out to be 12.6 when total students were divided by total number of teachers). Data for some districts was obviously incorrect (e.g., a reported per student expenditure of nine dollars), but the correct figure could not be recovered because errors were repeated across several columns of data. We also eliminated a number of districts because they were persistent outliers that ended up biasing the results—most of these were first or second year charter schools (these are treated as equivalent to school districts in the data set). The original N was 1102, including 61 charters. We eliminated 57 districts (including 24 charters) because of unreliable data, or because they were persistent outliers. This represents approximately 5 percent of the data set. Because of missing data on various variables, the actual Ns in our models ranged from 404 to 935. In most cases the actual N used to generate estimates represent between 80 and 90 percent of the original number of observations.

[5] Meier et al. (1999) also report negative coefficients when they model a purely linear relationship. The claim of a positive impact of minority teachers on school outputs is based on an analysis of threshold effects.

[6] We calculate the Herfindahl Index thus: 1 - Σ (District Enrollment / County Enrollment)^2. This results in a 0 to 1 index, where 0 indicates a single district monopolizes the entire enrollment in a county, and values close to 1 indicates several districts with equal-sized enrollments in the county.

TABLE 2
TEST VARIABLE PERFORMANCE WITH ALTERNATE OUTPUT MEASURES

Variable	TAAS Pass Rates	College Criterion	Attendance Rates	Completion Rate	Black Equity	Hispanic Equity
Percent Central Administration	Negative	Insignificant	Positive	Positive	Insignificant	Insignificant
Percent School Administration	Insignificant	Negative	Insignificant	Insignificant	Insignificant	Positive
Teacher/Professional Ratio	Insignificant	Negative	Positive	Positive	Insignificant	Negative
Percent Black Teachers	Negative	Negative	Negative	Insignificant	Insignificant	Insignificant
Percent Hispanic Teachers	Insignificant	Negative	Insignificant	Insignificant	Positive	Positive
Herfindahl Index	Insignificant	Insignificant	Insignificant	Insignificant	Insignificant	Insignificant
Instructional Expenditures	Insignificant	Negative	Positive	Positive	Insignificant	Insignificant

Cell entries indicate coefficient direction/significance of variable for given output measure when included in the fully-specified model reported in Table 1.

opportunities. The 10th grade TAAS is a required exit exam, the SAT/ACT helps determine access to higher education, so both have important socioeconomic consequences for the individual. Coleman (1974) argues that equality of educational opportunity can be operationally defined as the gap between ethnic groups on school outputs with these sorts of consequences. Accordingly we constructed an equity index consisting of a ratio of the percent of minority students passing the TAAS test to the percent of Anglos passing. A score of 1.0 indicates minority students are passing these exams at the same rates as Anglos, less than 1.0 indicates Anglos are passing at greater rates than minorities (there were only a handful of districts with scores higher than 1.0).

There is no direct measure of socialization in the data set, but there are a number of variables that have been used as proxies for socialization in other studies. We use a measure of attendance (percent of students attending on an average day) and a completion rate index (percent of the 9th grade cohort of 1994-95 who graduated, received a GED, or continued their education by enrolling in the 1998-99 school year). As dropouts and truants are much more likely to engage in socially deviant behavior, these sorts of measures are argued to represent a valid index of educational outcome in terms of inculcation into the basic values of good citizenship (Bryk, Lee, and Holland 1993; Coleman and Hoffer 1987; McNeal 1997).

Table 2 shows the performance of the test variables when the full model reported in Table 1 is used to predict the different dependent variables. Note that the inference for any given variable shifts according to the output measure employed as the dependent variable. For example, more top heavy districts—those with more staff resources in central administration—tend to have lower TAAS pass rates, but higher attendance and graduation rates. This makes sense from an economies-of-scope perspective. Larger administrative operations will have greater capacities to track and deter truancy and to run programs targeting at risk students. Success in these areas may keep at risk students in school, but these are also students who traditionally perform comparatively poorly on standardized tests so overall pass rates could easily fall. Similar tradeoffs are made clear elsewhere. Where significant, the variables measuring percent minority teachers are negatively related to overall test scores, but positively related to test score equity. Expenditures have no impact on equity, little (perhaps even negative) impact on test scores, but are positively related to attendance and graduation rates. Only the competition variable produced a uniform result across output measures, and that was to be uniformly insignificant. Any claims, and certainly any policy inference, taken from the results presented in Table 1 are put into a different context by the results in Table 2. The multidimensional nature of school goals means increasing the level of one input may boost one output while constraining another.

The second component of our second best solution was to explore coefficient variability under alternate, theoretically defensible, model specifications. Each of the test variables was systematically dropped and re-added to the model—in effect taking turns being sent off to the error term. We also experimented with including alternate measures in our "core variables" (e.g., average teacher salary). Finally, we tried several alternative ways to measure competition. These included using a simple concentration index (the percent of a county's enrollment in a single district) and the presence of a charter school. The minimum and maximum coefficient values for all test variables under these alternate specifications are reported in Table 3.

Of the 42 sets of Leamer bounds reported in Table 3, more than a quarter (12) are not bounded away from zero. This means in a significant minority of our alternate specifications the coefficient of a test variable switched signs. Under some specification scenarios these highly unstable

TABLE 3
LEAMER BOUNDS

Variable	TAAS Pass Rates	College Criterion	Attendance Rates	Completion Rate	Black Equity	Hispanic Equity
Percent Central Administration	–.815 : –.221	–1.1 : –.34	.05 : .13	.67 : .89	–.001 : .015 *	–.009 : –.006
Percent School Administration	–.26 : .01*	–1.5 : –1.2	–.03 : .03*	–.05 : .03*	–.005 : .008	.003 : .013
Teacher/Professional Ratio	.01 : .04	–.16 : –.12	.007 : .01	.04 : .06	–.002 : –.001	.001 : .001
Percent Black Teachers	–.29 : –.24	–.24 : –.12	–.02 : –.01	–.03 : .008*	–.003 : –.001	–.001 : .001*
Percent Hispanic Teachers	–.04 : .02*	–.15 : –.09	–.003 : .003*	.003 : .022	.001 : .002	.00 : .001
Herfindahl Index	.49 : –1.65*	2.5 :– .812*	.118 : –.13*	–.21 : –.88	.05 : .02	.03 : .005
Instructional Expenditures	.00 : .865	–1.5 : –.205	.124 : .198	.57 : .77	–.02 : –.003	.002 : 012

*Coefficient that changes direction based on model specification.

coefficients managed to clear the $p < .05$ threshold. Such models reported alone could be highly misleading about the relationships actually present in the data. The overall picture presented by Table 3, however, supports the basic inferences taken from Table 2. Percent staff in central administration, for example, was consistently a negative predictor of test scores and just as consistently a positive predictor of attendance and graduation rates. Expenditures were consistent, positive predictors of attendance, graduation rates, TAAS scores and Hispanic equity. They were consistent negative predictors of the college criterion variable and black equity measures (though in the latter, always very close to zero). The most unstable coefficients were the variable for percent in school administration (switching signs across four of the six dependent variables) and competition (switching signs across three of the dependant variable). Our alternate measures of competition did nothing to stabilize the latter result. We found it was possible to get the "right" sign on some of the competition variables in some of the models, but the only significant variables we could produce were in the "wrong" direction (as Bayesians might put it, the likelihood function changed the direction of our prior).[7] Across alternate specifications and differing output measures competition was highly unstable.

DISCUSSION

One might argue that our second best approach to education production functions for Texas school districts has simply replicated in miniature the entire problem we set out to investigate. We have presented a set of conflicting results using model specifications based on varying theoretical interests and arrived at no hard and fast conclusions—an apt description of the entire education production function literature. In contrast, we believe we have explained in no small measure why those contradictions exist and suggested a set of options for making production function analysis a more reliable platform for policy analysis and theory confirmation.

First, we believe using alternate output measures serves to highlight the limitations of theory. For example, education as been used as a laboratory to test frameworks like public choice and the theory of representative bureaucracy, neither of which is oriented toward a specific educational goal. Our results suggest more ethnically diverse teaching populations may help reduce test score gaps, but are not likely to do much in the way of boosting overall test scores (and may even do the opposite). Similarly, our findings suggest the neo-institutionalist arguments derived from public choice theory are only partially correct in suggesting that bureaucracies will constrain school outputs. Bureaucracy may constrain *some* outputs. There are reasons to expect it to boost other outputs. The second best test presented here suggests these theories are incomplete rather than incorrect in their explanations of education outputs. By demonstrating the limits of these frameworks and highlighting the economies of scope in education, a second best analysis sets a higher bar for theory confirmation and the claims attached to policy prescription.

Second, the deliberate attempt explore the fragility of coefficients limits the advance of normative creep. In one sense, the second best approach is what might be termed "closet Bayesianism"—it is designed to reveal subjective prior preferences even if these are unacknowledged by the researcher. Leamer bounds are a simple way to make transparent the ad hoc specification searches that underlie most published regression analyses. Such transparency makes it much harder to defend a preferred outcome if it is more an

[7] A table with all coefficients/Leamer bounds for all of the variables in all of models is posted athttp://www.unl.edu/s_rademacker/schools.htm.

artifact of particular specification choices rather than the consistent ending to the story told by the data. In some of our specification scenarios we could generate a positive coefficient for a competition variable just beyond the $p < .05$ criterion (i.e., $p < .10$) in a model that included a negative coefficient for expenditures. While such findings might be welcomed by certain policy agendas, in the context of this data set both findings are more accurately described as statistical artifacts or as relationships tied to single output measure.

Third, we would suggest the "messier" picture that in most cases is going to emerge from a second best approach could make empirical analyses more relevant to policy debates. An explicit motive of Leamer (1978, 1983) in proposing his simple form of sensitivity analysis was the recognition that consumers of empirical studies were becoming increasingly distrustful of them. Leamer observed that rather than adding a measure of scientific certainty to policy debates, empirical studies seemed in danger of becoming just another partisan political tool. Anyone reading the educational policy literature of the past 20 years may well conclude that danger has been realized. In contrast to many other studies, using the second best approach we found it virtually impossible to provide a clear "winner" in terms of theory, ideology or policy preference. We believe this is a key advantage of a second best analysis—in effect it helps clarify the available political choices rather than backing one over the other. For example, educational expenditures may not buy higher test scores. Yet, all else equal, Texas school districts that spend more on instruction seem to do a better job of keeping students in school. The acceptable equilibrium between dollars and achieving these two educational goals is a political choice that is not made any easier by this (or any other) empirical analysis. The second best approach, however, provides a clearer idea of the consequences and tradeoffs involved in those decisions. This, we believe, is an important contribution for empirical analysis to aspire to, and one that can be extended to policy analysis beyond the confines of education.

REFERENCES

Arnstine, Donald. 1995. *Democracy and the Arts of Schooling.* New York: State University Press of New York.

Becker, Gary S. 1993. *Human Capital,* 3rd ed. Chicago: University of Chicago Press.

Belfield, Clive R., and Henry M. Levin. 2002. "The Effects of Competition on Educational Outcomes: A Review of U.S. Evidence." National Center for the Study of Privatization in Education, Teachers College, Columbia University.

Bohte, John. 2001. "School Bureaucracy and School Performance at the Local Level." *Public Administration Review* 61: 92-99.

Bryk, Athony, Valerie Lee, and Peter Holland. 1993. *Catholic Schools and the Common Good.* Cambridge, MA: Cambridge University Press.

Burtless, Gary, ed. 1996. *Does Money Matter?: The Effect of School Resources on Student Achievement and Adult Success.* Washington, DC: Brookings Institution Press.

Chubb, John, and Terry Moe. 1988. "Politics, Markets and the Organization of Schools." *American Political Science Review* 82: 1065-1087.

_____. 1990. *Politics, Markets and America's Schools.* Washington, DC: Brookings Institution Press.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James J. McParland, Alexander M. Mood, Frederic D. Weinfield, and Robert L. York. 1966. *Equality of Educational Opportunity.* Washington, DC: U.S. Government Printing Office.

Coleman, James. 1974. "The Meaning of Equal Educational Opportunity." In LaMar P. Miller and Edmund W. Gordon, eds., *Equality of Educational Opportunity,* New York: AMS Press, Inc.

Coleman, James, and Thomas Hoffer. 1987. *Public and Private High Schools: The Impact of Communities.* New York: Basic Books.

Downs, George W., and Patrick D. Larkey. 1986. *The Search for Government Efficiency.* Philadelphia, PA: Temple University Press.

Easton, David. 1965. *A Framework for Political Analysis.* Englewood Cliffs, NJ: Prentice-Hall.

Fortune, Jim C. 1993. "Why Production Function Analysis is Irrelevant in Policy Deliberations Concerning Educational Funding Equity." *Educational Policy Analysis Archives.* Http://epaa.asu.edu/epaa/v1n11.html. Accessed March 12, 2001.

Frank, Kenneth. 2000. "Impact of a Confounding Variable on a Regression Coefficient." *Sociological Methods & Research* 29: 147-94.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52 (3): 647-74.

Glenn, David. 2001. "The Voucher Vortex." *Lingua Franca* 11 (4): 32-41

Hanushek, Eric A. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19: 141-64.

Henig, Jeffrey. 1994. *Rethinking School Choice: Limits of the Market Metaphor.* Princeton, NJ: Princeton University Press.

Hoxby, Caroline. M. 2000. "Does Competition Among Pubic Schools Benefit Students and Taxpayers?" *American Economic Review* 90: 1209-38.

Jencks, Christopher. 1998. "Racial Bias in Testing." In Christopher Jencks and Meridith Phillips, eds., *The Black-White Test Score Gap,* Washington, DC: Brookings

King, Gary. 1995. "Replication, Replication." *PS: Political Science & Politics* 28: 444-52.

Lankford, Hamilton, and James Wyckoff. 1992. "Primary and Secondary School Choice Among Public and Religious Alternatives." *Economics of Education Review* 11: 317-37.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley.

_____. 1983. "Let's Take the 'Con' Out of Econometrics." *American Economic Review* 73: 31-42.

Leamer, Edward E., and Herman Leonard. 1983. "Reporting the Fragility of Regression Estimates." *Review of Economics and Statistics* 65: 306-17.

Maranto, Robert, Scott Milliman, and Scott Stevens. 2000. "Does Private School Competition Harm Public Schools? Revisiting Smith and Meier's 'The Case Against School Choice.'" *Political Research Quarterly* 53: 177-92.

McNeal, R. 1997. "High School Dropouts: A Closer Examination of School Effects." *Social Science Quarterly* 78: 209-22.

Meier, Kenneth J. 1997. "Bureaucracy and Democracy: The Case for More Bureaucracy and Less Democracy." *Public Administration Review* 57 (3): 193-99.

Meier, Kenneth J., Warren S. Eller, Robert D. Wrinkle, and J. L. Polinard. 2001. "Zen and the Art of Policy Analysis: A Response to Nielsen and Wolf." *Journal of Politics* 63: 616-29.

Meier, Kenneth J., and Jeff Gill. 2000. *What Works: A New Approach to Program and Policy Analysis.* Boulder, CO: Westview Press.

Meier, Kenneth J., and Joseph Stewart, Jr. 1991. *The Politics of Hispanic Education.* Albany: State University of New York Press.

Meier, Kenneth J., Robert D. Wrinkle, and J. L. Polinard. 1999. "Representative Bureaucracy and Distributional Equity: Addressing the Hard Question." *Journal of Politics* 61 (November): 1025-39.

Miller, Trudi C. 1981. "Political and Mathematical Perspectives on Educational Equity." *American Political Science Review* 75: 319-33.

Monk, David H. 1992. "Education Productivity Research: An Update and Assessment of its Role in Education Finance Reform." *Educational Evaluation and Policy Analysis* 14 (4): 307-32.

Nielsen, Laura B., and Patrick J. Wolf. 2001. "Representative Bureaucracy and Harder Questions: A Response to Meier, Wrinkle, and Polinard." *Journal of Politics* 63: 598-615.

Owen, David. 1985. *None of the Above.* Boston, MA: Houghton Mifflin.

Pearl, Judea. 2000. *Causality: Models, Reasoning and Inference.* New York: Cambridge University Press.

Rebell, Michael. 1998. "Fiscal Equity Litigation and the Democratic Imperative." *Journal of Education Finance* 24: 23-50.

Rothstein, Richard. 1997. *What Do We Know About Declining (or Rising) Student Achievement?* Arlington, VA: Educational Research Service.

Schneider, Mark, Paul Teske, and Melissa Marschall. 2000. *Choosing Schools.* Princeton, NJ: Princeton University Press.

Smith, Kevin B., and Kenneth J. Meier. 1995. *The Case Against School Choice.* New York: Sharpe.

Treadway, Jack M. 1985. *Public Policymaking in the American States.* New York: Praeger.

Tyack, David, and Larry Cuban. 1995. *Tinkering Toward Utopia: A Century of Public School Reform.* Cambridge: Harvard University Press.

Verstegen, Deborah A., and Richard A. King. 1998. "The Relationship Between School Spending and Student Achievement: A Review and Analysis of 35 Years of Production Function Research." *Journal of Education Economics* 24: 243-62.

Weiher, Gregory. 2000. "Minority Student Achievement: Passive Representation and Social Context in Schools." *Journal of Politics* 62 (August): 886-95.

Wenger, Jennie. 2000. "What Do Schools Produce? Implications of Multiple Outputs in Education." *Contemporary Economic Policy* 18: 27-36.

Wilson, James Q. 1989. *Bureaucracy: What Government Agencies Do And Why They Do It.* New York: Basic Books.

Witte, John F. 2000. *The Market Approach to Education.* Princeton, NJ: Princeton University Press.

Wood, B. Dan, and Richard Waterman. 1994. *Bureaucratic Dynamics: The Role of Bureaucracy in a Democracy.* Boulder, CO: Westview Press.

Wrinkle, Robert D., Joseph Stewart, Jr., and J. L. Polinard. 1999. "Public School Quality, Private Schools and Race." *American Journal of Political Science* 43: 1248-53.

Zellner, A., J. Kmenta, and J. Dreze. 1966. "Specification and Estimation of Cobb-Douglas Production Function models." *Econometrica* 34 (4): 784-95.

Received: May 28, 2002
Accepted for Publication: July 16, 2002
ksmith1@unl.edu
jrademac@unlserve.unl.edu