THE USE OF MUTUALLY INTERDEPENDENT VS. MUTUALLY INDEPENDENT SCHOOL SYSTEM OUTPUTS IN ESTIMATING EDUCATION PRODUCTION FUNCTIONS*

Thomas G. Fox, The Pennsylvania State University

1. INTRODUCTION

During the past few years, school system "outcomes" (or outputs) such as "academic achievement" and "dropout prevention or 'holding power'" have been studied in relation to the school system resources available to "produce" these outputs. The functions relating school system inputs to outputs are known as education production functions in the literature of the economics of education. Though it is unequivocally accepted that very complex schooling processes "turn-out" more than one type of schooling or education output, previous empirical work in this area has statistically emphasized the mutual independence of education outputs while most authors verbally acknowledged output interdependence. This study builds on previous empirical work by estimating an interdependent output high school production function for a large-city high school system using simultaneous euqations and comparing results obtained with ordinary-least-squares estimates. The sensitivity of the statistical significance of coefficients obtained to the weighting method used is then considered.

The analysis is developed in three sections. Section 2 contains an overview of education production functions with interdependent outputs and the accompanying structural equations together with a discussion of alternative methods of weighting variables. Two-stage-least-squares (2SLS) regression estimates of the structural equations in the high school production function are presented in Section 3. In Section 4, ordinary-least-squares (OLS) regression analysis is applied to the structural equations with one output variable deleted to obtain naive estimates of reduced-form coefficients. Alternative estimates of the reduced-form coefficients are derived from 2SLS estimates of the coefficients in the structural equations. Estimated coefficients obtained are then comparatively analyzed. Conclusions are restated in the last section.

2. OVERVIEW

In a very general form, an education production function with two interdependent outputs can be written as

$$F(SI, PC, GE, AA, HP) = 0 \qquad (1)$$

where SI (school system resource inputs), PC(relevant personal pupil characteristics), and GE(general socioeconomic environment) are--in the short-run--given, predetermined, and therefore exogenous. The output variables AA (academic achievement) and HP (school holding power; i.e., 1 - dropout rate) are jointly determined as endogenous variables within the school system in interaction with the exogenous variables and with each other. Thus the education putput variables are hypothesized to be mutually interdependent.

Equation (1) contains two production functions, one for the production of academic achievement

$$AA = F_1(SI, PC, GE, HP) \qquad (2)$$

and one for the production of holding power --

$$HP = F_2(SI, PC, GE, AA). \qquad (3)$$

These are the structural equations for the education production function and are to be solved simultaneously.

Authors of the earlier single-equation studies [eq., 2,3,4] have recognized that since the size of each school affects the volume of school system inputs, input variables tend to be highly intercorrelated if measured as, say, total teacher man-years, when school enrollments vary from very small to very large. To reduce multicollinearity among the exogenous variables, it has been common practice to express all variables as averages per student, then re-insert school size by including enrollment or attendance as a separate variable. Thus variables usually include average achievement per student, median family income, average class size, among others.

Such a method of deflating "total school inputs" and "total school outputs" may or may not facilitate statistical estimation of underlying relationships among variables. In this study, each equation will be estimated in "average" form and in "total" form so that the statistical results can be compared.

Space does not permit a review of the numerous difficulties inherent in defining and measuring variables for education production functions. These problems have been carefully considered elsewhere [1,2,3]. The reader is forewarned, however, that data limitations and inadequate knowledge of "schooling technology" severely constrain the applicability of the education production function concept to all of the empirical studies that the author is aware of, including the present study. Measurement is imperfect, variables for pupil characteristics are usually unavailable, census data or attendance area or district-wide averages must often be used as "proxies" for pupil characteristics and for the general environment, (and hence empirically the PC's have been dropped from equations 1 and 2).

In addition to problems in specifying the proper variables, there are significant issues involved in specifying the mathematical form of the production functions. For the empirical work involved in this study, a log-linear education production function is assumed, as this type of function has proved highly useful in other production function studies. The structural equations assumed to represent the education production function are

$$AA' = N' \; X_3^{\alpha_3'} \; X_4^{\alpha_4'} \; X_5^{\alpha_5'} \; X_7^{\alpha_7'} \; X_8^{\alpha_8'} \; X_9^{\alpha_9'} \; X_{10}^{\alpha_{10}'} \; HP^{\lambda'} + b_{12}' X_{12} + b_{13}' X_{13}$$

*The author gratefully acknowledges computational services provided by the Computation Center, The Pennsylvania State University.

$$HP'' = N'' X_2^{\alpha_2''} X_4^{\alpha_4''} X_5^{\alpha_5''} X_6^{\alpha_6''} X_7^{7''} X_8^{8''} X_9^{9''} X_{10}^{10''} AA^{2''} + b_{12}'' X_{12} + b_{13}'' X_{13} \quad (5)$$

where

- X_2 = student employment, a proxy for ability of students to reduce the opportunity cost of a high school education in terms of foregone earnings
- X_3 = students planning on college attendance, a proxy for student preferences
- X_4 = high school attendance area income, a proxy for income of students' families
- X_5 = student class hours in vocational courses to adjust for curricula-mix differences among schools
- X_6 = years teaching experience of teachers, a proxy for the quality of man-years of teacher time
- X_7 = teacher man-years
- X_8 = auxiliary-service man-years (primarily teachers not in class-rooms such as librarians)
- X_9 = text and library book expenditures, a proxy for materials, supplies, other non-building educational capital inputs
- X_{10}= coded building age--weighted by attendance, a proxy for capital plant facilities of different ages and sizes
- X_{12}= dummy variable for racial composition of high school attendance area = 1 if less than 11% NW, otherwise 0, a proxy for racial mix of students
- X_{13}= dummy variable for racial composition of high school attendance area = 1 if 11% - 45% NW, otherwise 0, a proxy for racial mix of students
- AA = 11th grade reading achievement, a proxy for 9th-12th grade academic achievement
- HP = holding power of schools (1 minus drop-out rate for "average" weight equations, multiplied by number of students for "total" weight equations)
- N', N" = constant terms in equations

Date are for the 39 Chicago public high schools and are discussed at length in Burkhead-Fox-Holland [2].

The mathematical form assumed is log-linear in the parameters α_i, requiring all variables with α_i coefficients to have non-zero values before any production is forthcoming and has the well-known desirable economic features associated with Cobb-Douglas-type production functions. In simultaneous equation form each exogenous variable has both a direct affect <u>and</u> an indirect affect on <u>each</u> endogenous output variable: teacher man-years, for example, are expected to directly effect student achievement, and have an indirect affect operating through the impact of holding power on achievement as holding power is simultaneously affected by changes in teacher man-years. The total affect of each exogenous variable upon each endogenous variable, consists of the sum of its direct and indirect effects. The total impact of exogenous variables is found from reduced-form estimates of the structural equations. Hence any statistical estimating techniques which capture only the direct impact of exogenous input variables upon endogenous output variables can yield erroneous conclusions if estimated coefficients are interpreted as estimates of total effects.

3. 2SLS ESTIMATES

Estimates of the coefficients for the structural equations of the high school production function obtained by application of 2SLS techniques are displayed in Table 1, with standard errors reported in parentheses. The first pair of coefficient columns are for equation (4), the second pair pertain to equation (5). Headings titled "average units" are equations expressed in averages per student, per school, form; the headings "total units" are for totals by school.

Looking first at the structural equations for achievement, we observe that the weighting method affects the number of statistically significant coefficients, if a two-tailed t-test is applied and we require coefficients to be at least twice the size of their respective standard errors (significant coefficients are marked with asterisks). Two coefficients, one socioeconomic (X_4, income) and one for a school system input (X_8, auxiliary service man-years) are significant with "average" weights. Seven coefficients--three socioeconomic (X_4, income, and X_{12} and X_{13} for race), one for student preferences (X_3, students planning on college attendance), two school inputs (X_8, auxiliary service man-years and X_{10}, building age-attendance weighted) and the endogenous holding power variable (HP)--are significant in the "total form" structural equation for AA. The coefficients on the exogenous variables in the structural equations record only their direct affects upon achievement; all indirect affects operate through the other endogenous variable, holding power.

Interestingly, where coefficients are significant in <u>both</u> "average" and "total" form (X_4, X_8) they differ from each other by less than one standard error of either coefficient and hence are not statistically different.

The equation in "total" form explains almost 98 percent of the variance in total AA, whereas in "average" form, the same set of variables accounts for 93 percent of the variance in average AA. The endogenous variable HP is significant only in the "total" weighted structural equation for the production of achievement. Apparently, expressing all variables as "averages per students" to reduce multicollinearity among exogenous variables leads to a weaker explanatory equation with fewer individually significant coefficients.

The second pair of columns in Table 1 pertain to equation (5). The structural equation for holding power in "total" form explains almost 99 percent of the variance in total holding power; when estimated in "average" form, the same set of variables explain only 60 per cent of the variance in holding power. Excluding constant terms, the "total" form yields 6 coefficients whose values are more than twice the size of their standard errors, the "average" weighted equation has 4 significant coefficients. In the two instances where both "average" and "total" coefficients are significant (AA, X_7) they differ in magnitude by more than two standard errors for teacher-man-years but by less than two standard errors for achievement (when only the standard error for the "total" weight coefficient is used), and their signs are all positive.

TABLE 1
TWO-STAGE-LEAST-SQUARES ESTIMATES OF STRUCTURAL EQUATIONS
FOR A HIGH SCHOOL SYSTEM PRODUCTION FUNCTION

		Achievement		Holding Power	
		Average Units	Total Units	Average Units	Total Units
AA	Achievement-Stanines (Endogenous)			.444* (.010)	.381* (.146)
HP	Holding Power (Endogenous)	4.022 (3.544)	.171* (.072)		
X_{12}	Race=1 if < 11% NW, Otherwise 0	.001 (.031)	.081* (.021)	- .003 (.007)	- .038* (.011)
X_{13}	Race=1 if 11-45% NW, Otherwise 0	- .024 (.825)	.052* (.026)	.008 (.009)	- .015 (.013)
X_2	Student Employment			- .024 (.034)	.112* (.054)
X_3	Students Planning on College Attendance	.082 (.171)	.227* (.107)		
X_4	Area Income ($100)	.526* (.086)	.489* (.172)	- .228* (.061)	- .015 (.097)
X_5	Vocational Class Student Hours	- .067 (.080)	- .019 (.092)	.018 (.023)	.099* (.027)
X_6	Years Teaching Experience			- .001 (.005)	- .012 (.012)
X_7	Teacher Man-years per 100 Students	- .114 (.061)	.025 (.272)	.053* (.026)	.326* (.075)
X_8	Auxiliary Service Man-years per 100 Students	.223* (.068)	.159* (.073)	- .087* (.027)	- .046 (.044)
X_9	Text and Library Book Expenditures ($)	.003 (.010)	- .031 (.067)	.007 (.018)	.065* (.025)
X_{10}	Building Age-Weighted by Attendance	.014 (.025)	.042* (.021)	- .007 (.011)	- .001 (.026)
	Constant	-7.099 (7.008)	.056 (1.037)	1.605* (.117)	1.359* (.144)
	R^2	.927	.978	.598	.987

Though it would be interesting to study and analyze the indivdual variables at greater length, deducing certain plausible educational implications, we continue to concentrate on comparisons. As previous education production functions have used single equation estimating models rather than simultaneous-equation models, we now turn to the single-equation comparisons.

4. REDUCED-FORM ESTIMATES

In this section, alternative techniques for assessing the total statistical impact of exogenous variables upon endogenous output variables are considered using achievement, as the endogenous variable.

Suppose first, that some researcher specifies his education production functions exactly as we have specified the structural equations (4) and (5). Suppose, further, that for arbitrary or a priori reasons he concludes that the education outcomes are mutually independent and hence not jointly determined. He therefore deletes $HP^{\lambda'}$ from (4) and $AA^{\lambda''}$ from (5): the resulting single-equation estimating models contain only exogenous variables hypothesized to affect only the single dependent variable in each equation. This version of statistical estimating models, deliberately misspecified (from the context of the previous section) is labled as "simple OLS" in Table 2.

If the analyst recognizes that exogenous variables have both direct and indirect effects on endogenous variables, he will no doubt be interested in estimating their total effect on each endogenous variable. This is found by eliminating all but one endogenous variable from the production function in order to obtain the reduced-form equation of the structural education production function. Two different statistical procedures can be employed to estimate these reduced-form coefficients, OLS and 2SLS regression techniques. In estimating reduced-form coefficients, it is widely recognized in the econometrics literature that OLS yields biased and inconsistent results so 2SLS is preferred [5, p. 189]. In many economic empirical studies, actual estimates of coefficients by either OLS or 2SLS techniques frequently yield similar results, though 2SLS is theoretically preferrable to OLS. Both techniques are utilized herein. Previous empirical studies of education production functions where various education output variables, one at a time, have been regressed against a common set of independent variables

TABLE 2
ACHIEVEMENT: COMPARISONS OF ALTERNATIVE ESTIMATES OF SINGLE EQUATION HIGH SCHOOL SYSTEM PRODUCTION FUNCTIONS

		Average Unit Weights			Total Unit Weights		
		Simple[a] OLS	Reduced-Form Coefficients OLS	Reduced-Form Coefficients 2 SLS	Simple OLS	Reduced-Form Coefficients OLS	Reduced-Form Coefficients 2 SLS
X_{12}	Race=1 if < 11% NW, Otherwise 0	.032 (.019)	-.010 (.014)	.014	.083* (.020)	.009 (.013)	.214
X_{13}	Race=1 if 11-45% NW, Otherwise 0	.001 (.059)	.001* (.000)	-.010	.055* (.022)	.001 (.000)	.134
X_2	Student Employment		.048 (.779)	.123		.101 (.055)	.054
X_3	Students Planning on College Attendance	-.112* (.029)	-.065* (.025)	-.104	.250* (.051)	.083* (.032)	.649
X_4	Area Income	.553* (.096)	.467* (.073)	.498	.526* (.075)	.492* (.061)	1.389
X_5	Vocational Class Student Hours	.002 (.053)	-.023 (.047)	-.006	.001 (.071)	-.041 (.030)	-.006
X_6	Years Teaching Experience		.006 (.016)	.005		.007 (.015)	.006
X_7	Teacher Man-years	-.141* (.066)	-.332* (.062)	-.126	.083 (.112)	-.020 (.090)	.231
X_8	Auxiliary Service Man-years	.202* (.076)	.068 (.066)	.162	.168* (.066)	.046 (.055)	.431
X_9	Text and Library Book Expenditures ($)	-.013 (.047)	-.027 (.716)	-.032	-.017 (.035)	-.020 (.030)	-.057
X_{10}	Building Age-Weighted by Attendance	.013 (.028)	-.006 (.023)	.018	.043 (.025)	.015 (.020)	.120
	Constant	.858* (.224)	1.122* (.183)	.819	.303 (.179)	1.608* (.241)	.823
	R^2	.901	.940		.976	.986	

Notes: [a]OLS estimates of structural equations for achievement after deleting high school holding power as a variable.

using OLS (such as Burkhead-Fox-Holland) can be properly interpreted as (naive) estimates of the reduced-form of structural education production functions even though this was not the original intent of the authors.

Turning to Table 2, note that the deliberately misspecified simple OLS equation yields as many significant coefficients for exogenous variables (four) as the OLS model, using average unit weights, and simple OLS yields twice as many, when total unit weights are utilized, yet the OLS always explains more of the variance in academic achievement. The number of statistically significant coefficients is quite sensitive to the specification of variables in the estimating equations. (As existing tests of significance for 2 SLS estimated reduced-form coefficients, especially from small samples, are questionable, these are not presented: 2 SLS reduced-form coefficients will be compared with significant simple OLS and OLS coefficients.) The signs of the coefficients are sensitive to the model used in only one case: where average unit weights are used, the race variable X_{13} has a positive sign for the simple OLS and OLS variants, a negative sign for 2 SLS variant. Values for some coefficients are similar (e.g., average weight, X_4 coefficients are within 2 standard errors of each other); yet differ dramatically when the weighting system changes (e.g., X_4). Similar conclusions emerge from analysis of reduced-form equations for holding power, not presented herein to conserve space.

5. CONCLUSIONS

We find the empirical results obtained for the high school production function to be highly sensitive to the estimation techniques employed and variable weights utilized. If these characteristics are also embodied in other education production function studies, and they probably are, then potential operating policy conclusions should not be based on the empirical work, unless a careful sensitivity analysis has been developed which clearly delimits the applicability of the statistical models.

In this study, the 2 SLS "total" variant appears to provide the best statistical results.

In the estimation of the structural education production functions at least with this set of data, variables expressed in "total" weights yield equations with higher coefficients of multiple determination and a larger number of statistically significant coefficients than obtained when structural equations use "average" weighted variables. Further, the fact that three of the four coefficients on endogenous variables are statistically significant supports the hypothesis that the educational outputs of academic achievement and holding power are mutually interdependent.

Turning to the single equation models, we know that, theoretically, the simple OLS variant is improperly specified, given the production function hypothesized, yet it often yields more significant coefficients for exogenous variables than does OLS. Hence empirical results are frequently sensitive to specifications of variables in the production functions, when single equation estimating techniques are employed. Also, it is well-known that OLS gives a biased and inconsistent estimate of reduced-form coefficients: when compared with reduced-form coefficients derived from 2 SLS estimates, OLS estimated reduced-form coefficients are frequently quite different. The total statistical impact of several of the exogenous variables is sensitive to both the weighting method used and the statistical technique used. The 2 SLS method has the advantage of explicitly showing interdependence among the endogenous variables. Much more work on education production functions is needed before users of these studies can be sure that empirical results are not simple statistical artifacts.

SELECTED REFERENCES

[1] Bowles, Samuel, "Towards an Education Production Function," Education, Income and Human Capital, W. Lee Hansen, ed., Studies in Income and Wealth, Vol. 35, National Bureau of Economic Research, (New York: Columbia University Press, 1970), pp. 11-61.

[2] Burkhead, Jesse, Thomas Fox, and John Holland, Input and Output in Large-City High Schools, Syracuse, N.Y.: Syracuse University Press, 1967.

[3] Kiesling, Herbert J., "Measuring a Local Government Service: A Study of School Districts in New York State," Review of Economics and Statistics (1967), 356-67.

[4] ________________, "The Relationship of School Inputs to Public School Performance in New York State," The RAND Corporation, Santa Monica, California (1969).

[5] Walters, A.A., An Introduction to Econometrics, New York: W.W. Norton & Company, Inc., 1970.