Estimation of an Education Production Function under Random Assignment with Selection

By ELEANOR JAWON CHOI AND HYUNGSIK ROGER MOON AND GEERT RIDDER *

This paper estimates an education production function using data on the College Scholastic Ability Test (CSAT) score and high school characteristics from Seoul, Korea.[1] A unique institutional feature of the high school system in Seoul is that on entering high school students are randomly assigned to schools within each school district. The main contribution of our study is to derive a school production function by aggregating the individuals' potential outcome functions that depend on observed and unobserved school inputs interacted with heterogeneous and unobserved individual abilities. The school production function derived under random assignment and under the assumption that there are no cohort effects has three unique features that have not been considered in previous studies. First, its average (over students) coefficients on school inputs do not differ by school or over time, but by district. This is a consequence of the endogenous sorting of students between districts[2] combined with the random assignment to schools within districts. Second, it allows unobserved school effects to be potentially correlated with observed ones. Third, the weighted average of the district-specific school input effects with weights equal to the fraction of the population in the districts is equal to the average partial effect (APE) of school inputs on individual academic achievement. To estimate the school production function coefficients, we first obtain district-specific coefficients using the fixed effect estimation method in school level panel data for each district and compute the weighted average described above. The empirical findings are (i) the school production function coefficients do differ between districts, which may be due to potentially endogenous sorting of students or unobserved differences in district characteristics, (ii) our estimate of the single-sex school effect is much larger than that found in previous studies most of which assumed constant school input coefficients across districts and did not consider school fixed effects.

I. Background and Data

The education policy in Korea over the past four decades greatly emphasized equal educational opportunity. In accordance with the policy emphasis, the High School Equalization Policy (HSEP) was adopted in Seoul in 1974. The HSEP aimed to provide students with a uniform learning environment and to close the achievement gap across schools by minimizing across-school variation in student quality, teacher quality, and school facilities and curriculum.[3][4] Under the HSEP, students were randomly assigned to academic high schools within school districts where they met residency requirements.[5] The

* Choi: College of Economics and Finance, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul 133-791, Korea. Email: choiej@hanyang.ac.kr. Moon (corresponding author): Department of Economics, University of Southern California, KAP 300, Los Angeles, CA 90089-0253. School of Economics, Yonsei University, Seoul, Korea. Email: moonr@usc.edu. Ridder: Department of Economics, University of Southern California, KAP 300, Los Angeles, CA 90089-0253. Email: ridder@usc.edu. We thank Jaesung Choi, Cheolsung Park, Jinyoung Kim, Seik Kim, David Lee, Nayoung Lee and Hongliang Zhang for helpful conversations; Korea Education and Research Information Service (KERIS) and the Ministry of Education of the Republic of Korea for providing data; and Yoonju Jung for research assistance. Choi gratefully acknowledges funding from Hanyang University (HY-201300000000418-N). Moon appreciates financial support from the USC Lusk Center for Real Estate.

[1] Meghir and Rivkin (2011) provides an overview of the literature on education production. See references therein. Park, Behrman and Choi (2012) uses the same data as in this study to investigate the effect of single sex education on student achievement.

[2] For more information on school choice and residential sorting, see, for example, Bayer, Ferreira and McMillan (2007).

[3] The strong emphasis on equal treatment in education policy has been maintained until 2009. The policy focus has shifted from uniformity to diversity afterward. Policymakers started to encourage competition among schools in 2010.

[4] For more information on the HSEP and its impacts, see, for example, Kim, Lee and Lee (2008).

[5] The student assignment lottery covered academic high schools in ten school districts, including Districts 1-4, 6-11. High schools excluded from the random assignment were vocational

TABLE 1—SUMMARY STATISTICS FOR BOYS

	District 3	District 4	District 6	District 7	District 8	District 9
Average Korean CSAT score	91.9	97.6	98.2	98.4	102.8	94.2
	[3.7]	[3.3]	[2.1]	[4.7]	[3.1]	[2.4]
Percentage of single-sex schools	28.6	35.3	35.7	52.9	47.4	41.7
Percentage of private schools	21.4	41.2	57.1	70.6	52.6	50.0
Age of school in 2008 (in years)	25.1	19.1	42.1	27.9	40.0	32.4
	[14.5]	[19.0]	[35.5]	[23.6]	[28.4]	[15.8]
Senior class size	35.9	35.0	36.3	35.1	34.3	34.7
	[2.4]	[1.9]	[2.7]	[2.6]	[2.5]	[2.2]
Percentage of students receiving	10.5	7.1	5.7	7.6	3.9	11.6
lunch support	[4.7]	[3.9]	[2.2]	[4.4]	[2.9]	[3.1]
Annual development fund spending	30.6	26.1	48.5	31.2	94.9	48.2
per student (in 1000 KRW)	[24.8]	[39.3]	[54.0]	[26.6]	[96.6]	[60.9]
Percentage of female teachers	49.7	44.6	37.5	31.2	39.1	38.8
	[18.0]	[13.3]	[17.0]	[19.4]	[19.4]	[23.0]
Number of male seniors	297.6	316.7	417.2	347.4	344.7	282.8
per school	[114.9]	[175.9]	[143.0]	[156.8]	[153.5]	[140.9]
Number of male CSAT takers	267.6	297.2	380.0	326.0	314.2	257.6
per school	[101.3]	[170.1]	[130.3]	[152.2]	[138.4]	[130.7]
Number of high schools	14	17	14	17	19	12

Note: All variables are for the school level. Standard deviations in brackets. 1000 KRW is worth approximately 1 USD.

random assignment made the distribution of student ability similar among the schools within a district. Thus, students living in the same district had similar peers. When students and their families moved to another school district, the students were reassigned randomly to a school in the new district.

There are 55 coed, 34 all-girls, and 38 all-boys high schools, which are either public or private, in our data.[6] Until choice-based assignment was introduced in 2010, academic high schools were subject to the lottery-based assignment regardless of their type – coed vs. single-sex or public vs. private. Unlike in the US and many other countries, private academic high schools were not much different from public academic high schools in educational environment, school curriculum, government subsidy, teacher quality, and even school tuition.

We use data on the CSAT scores and high school characteristics obtained from the Korean Ministry of Education and KERIS. We link the individual level test score data and the high school characteristics using school names.[7] The CSAT is the standardized test for college admissions in Korea. This test is developed, published, administered, and scored by the Korean government. The CSAT score on Korean is the main educational outcome in this study.[8] The scores were standardized to have a mean of 100 points and a standard deviation of 20 points. The test is offered once a year in November and is taken by about 600,000 individuals including high school seniors, high school graduates, and GED holders. The CSAT score together with the high school GPA are the most important factors that determine, whether a student is admitted to some college and to which college.

We restrict our analysis to high school seniors in 2008 and 2009,[9] who were randomly assigned to academic high schools within school districts 3, 4, 6-9 of Seoul. All academic high schools within each of the six districts participated in the

high schools, selective high schools specialized in science, foreign languages, art, or physical education, and academic high schools near the city center – mostly in District 5 and some in Districts 1, 2, 10, and 11.

[6] Single-sex schools tend to be older and are more likely to be private. This is partly because in the past high schools started as single-sex schools. The government has increased the number of coed schools by requiring since 1998 that all newly-opened public schools are coed.

[7] Our data cover the entire population of CSAT takers and high schools in Korea, but contain no individual characteristics other than gender, whether the person is a high school student, and which high school the person attends.

[8] The CSAT consists of five major sections: Korean, Math, English, Sciences/Social Studies/Vocational Education, and Second Foreign Languages. The results for English and Math scores are not much different from the results for Korean scores.

[9] School characteristics are available from 2008 and the HSEP was effectively abolished in 2010.

lottery-based student assignment. For the 2008 and 2009 cohorts of seniors, the assignment was conducted in February 2006 and 2007, respectively.[10] The analysis sample covers about 60 percent of CSAT takers in Seoul – 50,809 students in 2008 and 58,905 in 2009. Table 1 shows means and standard deviations of school level variables that are included in our empirical specifications. We focus on boys here and numbers for girls are shown in the online appendix.

II. Econometric Framework

A. The Individual Potential Outcome

We consider the following *potential outcome* of individual $i \in \mathcal{I}$ at school $s(d) \in \mathcal{S}_d$ of district $d \in \mathcal{D}$ in year $t \in \mathcal{T}$. We assume a linear education production function with heterogeneous coefficients:

$$Y_i(s(d), d, t) = \mathbf{K}'_{s(d)}\boldsymbol{\alpha}_i + \mathbf{L}'_{s(d),t}\boldsymbol{\beta}_i + \upsilon_{s(d)}\omega_i + u_{s(d),t}\xi_i + c_d\eta_i. \tag{1}$$

The (potential) outcome $Y_i(s(d), d, t)$ is the (potential) CSAT score of student i if he attends school $s(d)$ of district d in year t. The variables $\mathbf{K}_{s(d)}$ and $\mathbf{L}_{s(d),t}$ denote time-invariant and time-varying school inputs, respectively. The variables $\upsilon_{s(d)}$ and $u_{s(d),t}$, respectively, represent the unobserved time-invariant school inputs and unobserved time-varying school inputs. The variable c_d represents (unobserved) district characteristics. The coefficients $(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \omega_i, \xi_i, \eta_i)$ represent heterogeneous individual responses to the school inputs (observed and unobserved) and the unobserved district characteristics. The specification of the potential outcome model for $Y_i(s(d), d, t)$ assumes that the potential outcome is determined by the interaction of the school level inputs and the district characteristics $(\mathbf{K}_{s(d)}, \mathbf{L}_{s(d),t}, \upsilon_{s(d)}, u_{s(d),t}, c_d)$ (observed and unobserved) and the individual (heterogeneous) coefficients $(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \omega_i, \xi_i, \eta_i)$.[11]

Suppose that S_i denotes the school that individual i attends, T_i denotes the senior year of individual i, and D_i is the district where individual i chose to live. The observed outcome, i.e. the CSAT score, of individual i is

$$Y_i = Y_i(S_i, D_i, T_i).$$

Note that the outcomes (Y_i, S_i, D_i, T_i) are observed at the individual level, and $(\mathbf{K}_{s(d)}, \mathbf{L}_{s(d),t})$ at the school level.

The parameters of interest are the APE of the school inputs of interest $\mathbf{K}_{s(d)}$ and $\mathbf{L}_{s(d),t}$:[12]

$$\boldsymbol{\alpha} = \mathbb{E}[\boldsymbol{\alpha}_i] \text{ and } \boldsymbol{\beta} = \mathbb{E}[\boldsymbol{\beta}_i].$$

B. School Production Function

The school production function is the aggregate of the individual outcome functions and depends on the school level inputs. The aggregation is done under the following two key assumptions.

ASSUMPTION 1: *We assume that for all* $(s(d), d, t)$,
$\mathbb{E}[(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \omega_i, \xi_i, \eta_i)|S_i = s(d), D_i = d, T_i = t]$
$= \mathbb{E}[(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \omega_i, \xi_i, \eta_i)|D_i = d, T_i = t]$.

ASSUMPTION 2: *We assume that for all* t,
$\mathbb{E}[(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \omega_i, \xi_i, \eta_i)|D_i = d, T_i = t]$
$= \mathbb{E}[(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \omega_i, \xi_i, \eta_i)|D_i = d]$.

Assumption 1 follows from the random assignment of students within school districts. Assumption 2 assumes that the district average of student abilities does not change over time. This assumption is justified if the distribution of student abilities and the district choice selection does not change across cohorts. Given that our data covers two consecutive years, Assumption 2 is reasonable.

The average input effects for students in district d are denoted by

$$(\boldsymbol{\alpha}_d, \boldsymbol{\beta}_d, \omega_d, \xi_d, \eta_d) = \mathbb{E}[(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \omega_i, \xi_i, \eta_i)|D_i = d].$$

If the individual district choice is independent of the individual input effect, then $\boldsymbol{\alpha} = \boldsymbol{\alpha}_d$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_d$. In the case we study, however, the average productivity of school inputs may differ by school district because students were likely to be sorted endogenously across districts. By

[10] The school year begins in early March and ends in mid February in Korea.

[11] In the potential outcome function (1), we do not include a time effect because the CSAT scores are normalized.

[12] $\mathbb{E}[\cdot]$ denotes the population average of individuals.

TABLE 2—SCHOOL INPUT EFFECTS ON KOREAN CSAT SCORES FOR BOYS

	(1) District 3	(2) District 4	(3) District 6	(4) District 7	(5) District 8	(6) District 9	(7) APE
Single-sex	9.78	0.89	1.28	8.46	5.15	2.49	4.64
	(2.31)***	(1.77)	(0.99)	(3.02)**	(1.45)***	(2.05)	(0.84)***
Senior class size	-0.46	-0.13	-0.34	0.11	-0.25	-0.01	-0.18
	(0.41)	(0.11)	(0.11)***	(0.19)	(0.13)*	(0.24)	(0.08)**

Note: *, **, and *** indicate significance at 10 percent, 5 percent, and 1 percent level, respectively. Robust standard errors in parentheses. Standard errors clustered in school level for coefficients on time-varying regressors. See the text for the list of time-varying and time-invariant control variables.

allowing the average productivity to differ between districts, we explicitly take into account the potentially endogenous district selection.

We define $Y_{s(d),t}$ as the average test score of school s in district d in year t. Under Assumptions 1 and 2, the average test score of school s can be expressed as

$$\begin{aligned} Y_{s(d),t} &= \mathbb{E}\left[Y_i | S_i = s(d), D_i = d, T_i = t\right] \\ &= \mathbf{K}'_{s(d)}\boldsymbol{\alpha}_d + \mathbf{L}'_{s(d),t}\boldsymbol{\beta}_d \\ &\qquad + \upsilon_{s(d)}\omega_d + u_{s(d),t}\xi_d + c_d\eta_d. \end{aligned}$$

For notational convenience, we will also use the simplified subscripts $Y_{s,d,t} = Y_{s(d),t}$, $\mathbf{K}_{s,d} = \mathbf{K}_{s(d)}$, $\mathbf{L}_{s,d,t} = \mathbf{L}_{s(d),t}$, $V_{s,d} = \upsilon_{s(d)}\omega_d$, $U_{s,d,t} = u_{s(d),t}\xi_d$, and $C_d = c_d\eta_d$. Then, we can write the average outcome of school s in district d and year t as a function of school inputs and district characteristics:

$$(2) \quad \begin{aligned} Y_{s,d,t} &= \mathbf{K}'_{s,d}\boldsymbol{\alpha}_d + \mathbf{L}'_{s,d,t}\boldsymbol{\beta}_d \\ &\qquad + V_{s,d} + U_{s,d,t} + C_d, \end{aligned}$$

which yields the school production function.

Note that we derive the school production function by aggregating the individual outcomes. This procedure is similar to the derivation of the market demand function as an aggregation over individual choices (Berry, Levinsohn and Pakes (1995)). The school production function (2) has the unique feature that the coefficients $\boldsymbol{\alpha}_d$ and $\boldsymbol{\beta}_d$ of the observed school inputs are district specific, but constant across schools within each district and over time. The random assignment of students within districts and the assumption of no cohort effects are key for the constant productivity of school inputs within a district. Notice that if there is no individual heterogeneity in the potential outcome, which is a very strong restriction, it follows that $\boldsymbol{\alpha}_d = \boldsymbol{\alpha}$ and $\boldsymbol{\beta}_d = \boldsymbol{\beta}$. Under self-selection of schools and individual heterogeneity, the school production function is a correlated random coefficient model, and the identification of the school input coefficients $\left(\boldsymbol{\alpha}_{s(d)}, \boldsymbol{\beta}_{s(d)}\right)$ using school level data becomes challenging. In our setup, district specific coefficients $\left(\boldsymbol{\alpha}_d, \boldsymbol{\beta}_d\right)$ are district averages of $\left(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i\right)$.

C. Estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

For identification, we assume that $\mathbb{E}\left[U_{s,d,t} | \left\{\mathbf{L}_{s,d,t} : t \in \mathcal{T}\right\}\right] = 0$, $\mathbb{E}\left[\sum_{t\in\mathcal{T}} U_{s,d,t}/T | \mathbf{K}_{s,d}\right] = 0$, and $\mathbb{E}\left[V_{s,d} | \mathbf{K}_{s,d}\right] = \mathbb{E}\left[V_{s,d}\right]$. The usual identification assumptions imply strict exogeneity of $\mathbf{L}_{s,d,t}$ with respect to time-varying unobserved school effects, $U_{s,d,t}$, and exogeneity of $\mathbf{K}_{s,d}$ with respect not only to the time average of $U_{s,d,t}$ but also to $V_{s,d}$.

The APE parameters of interests are

$$\boldsymbol{\alpha} = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\alpha}_i | D_i = d\right]\right] = \sum_{d\in\mathcal{D}} \boldsymbol{\alpha}_d \mathbb{P}\left(D_i = d\right),$$

$$\boldsymbol{\beta} = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\beta}_i | D_i = d\right]\right] = \sum_{d\in\mathcal{D}} \boldsymbol{\beta}_d \mathbb{P}\left(D_i = d\right).$$

We can estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by taking the averages of estimated $\boldsymbol{\alpha}_d$ and $\boldsymbol{\beta}_d$ weighted by the district choice probabilities:

$$\hat{\boldsymbol{\alpha}} = \sum_{d\in\mathcal{D}} \hat{\boldsymbol{\alpha}}_d \frac{N_d}{N} \text{ and } \hat{\boldsymbol{\beta}} = \sum_{d\in\mathcal{D}} \hat{\boldsymbol{\beta}}_d \frac{N_d}{N},$$

where N_d is the number of students in district d and N is the total number of students in Seoul.

In view that the school production function (2) takes a panel linear regression form within each district, we can obtain a within estimator of

β_d using fixed effect estimation district by district:[13]

$$\hat{\beta}_d = \left[\sum_{s\in\mathcal{S}_d}\sum_{t\in\mathcal{T}}\left[\mathbf{L}_{s,d,t}-\overline{\mathbf{L}}_{s,d,\bullet}\right]\left[\mathbf{L}_{s,d,t}-\overline{\mathbf{L}}_{s,d,\bullet}\right]'\right]^{-1} \times \sum_{s\in\mathcal{S}_d}\sum_{t\in\mathcal{T}}\left[\mathbf{L}_{s,d,t}-\overline{\mathbf{L}}_{s,d,\bullet}\right]\left[Y_{s,d,t}-\overline{Y}_{s,d,\bullet}\right].$$

Then, α_d can be estimated as follows:[14]

$$\hat{\alpha}_d = \left[\sum_{s\in\mathcal{S}_d}\left[\mathbf{K}_{s,d}-\overline{\mathbf{K}}_{\bullet,d}\right]\left[\mathbf{K}_{s,d}-\overline{\mathbf{K}}_{\bullet,d}\right]'\right]^{-1} \times\sum_{s\in\mathcal{S}_d}\left[\mathbf{K}_{s,d}-\overline{\mathbf{K}}_{\bullet,d}\right]\left[\frac{1}{T}\sum_{t\in\mathcal{T}}\left[Y_{s,d,t}-\mathbf{L}'_{s,d,t}\hat{\beta}_d\right]\right].$$

III. Results and Discussion

Table 2 presents the estimated school input effects, especially the effect of single-sex education and the senior class size for boys.[15] Regressions also include other time-varying and time-invariant covariates that serve as control variables and are possibly correlated with unobserved school characteristics. Time-varying controls include the fraction of students receiving free or reduced price lunch, annual development fund spending per student, and the fraction of female teachers. Time-invariant controls include a private school indicator, age of the school in 2008, and the interaction between the two.

From columns (1)-(6), we observe that single-sex education effects vary substantially across school districts from no effect in District 4 to a positive effect as large as half a standard deviation in District 3. The class size effect is near zero (or insignificant) and negative in all districts but District 7. The heterogeneous effects imply that endogenous sorting of individuals across districts may play an important role. To understand the mechanism of sorting, we would need more information on individual characteristics from which we could infer how school characteristics interact with individual preference and productivity. The estimated APE of school inputs are shown in column (7). We use the number of CSAT takers in each district to construct the weighted average.[16] Compared to Park, Behrman and Choi (2012) who used the same data but a different model specification,[17] our APE estimates are qualitatively similar but quantitatively different – the effect of single-sex education is much larger.

Our findings suggest that it is important to take district heterogeneity due to sorting between districts and unobserved school characteristics into account when estimating the average effect of school inputs on test score.

REFERENCES

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. “A Unified Framework for Measuring Preferences for Schools and Neighborhoods.” *Journal of Political Economy*, 115(4): pp. 588–638.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. “Automobile Prices in Market Equilibrium.” *Econometrica*, 63(4): pp. 841–890.

Kim, Taejong, Ju-Ho Lee, and Young Lee. 2008. “Mixing versus sorting in schooling: Evidence from the equalization policy in South Korea.” *Economics of Education Review*, 27(6): 697 – 711.

Meghir, Costas, and Steven Rivkin. 2011. “Econometric Methods for Research in Education.” In *Handbook of the Economics of Education*. Vol. 3, , ed. Stephen Machin Eric A. Hanushek and Ludger Woessmann, Chapter 1, 1 – 87. Elsevier.

Park, Hyunjoon, Jere R. Behrman, and Jaesung Choi. 2012. “Causal Effects of Single-Sex Schools on College Entrance Exams and College Attendance: Random Assignment in Seoul High Schools.” *Demography*, 1–23.

[13] We define school level averages of $\mathbf{L}_{s,d,t}$ as $\overline{\mathbf{L}}_{s,d,\bullet} = \sum_{t\in\mathcal{T}}\mathbf{L}_{s,d,t}/T$. $\overline{Y}_{s,d,\bullet}$ is defined in the same manner.

[14] Note that $\overline{\mathbf{K}}_{\bullet,d} = \sum_{s\in\mathcal{S}_d}\mathbf{K}_{s,d}/N_S(d)$, where $N_S(d)$ is the number of schools in district d.

[15] Results for girls are in the online appendix.

[16] The APE estimates change little when we use the number of seniors or the cohort size at random assignment as weights.

[17] They assume that school and district effects are random effects and that school input effects are constant across schools and districts.