8-2010

Migration and Child Educational Production: Aggregated vs. Disaggregated Resource Modeling

Guo Li
Clemson University, guol@clemson.edu

MIGRATION AND CHILD EDUCATIONAL PRODUCTION: AGGREGATED VS. DISAGGREGATED RESOURCE MODELING

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
of Doctor of Philosophy in
Economics

---

by
Guo Li
August 2010

Accepted by:
Dr. Mroz, Thomas A., Committee Chair
Dr. Maloney, Michael T.
Dr. Simon, Curtis J.
Dr. Warner, John T.

ABSTRACT

This paper studies the sensitivity of estimates on various assumptions about aggregation in modeling the school's effect in child educational production. By controlling for the endogeneity of school qualities in the production function, we evaluate the performance of a "correct" aggregation educational production model versus simple aggregation educational production model in predicting the school resources' effect on academic outcome. Monte Carlo simulations on different modeling specifications shows that simple aggregation of school resources over a geographic area causes serious specification errors, and thus generates biased estimates for the marginal contribution of the school resources to test scores. The two aggregation models are empirically estimated, and we find that having heterogeneity control in the production function reduces the estimated effect of school characteristics on test score. We also find that the "Correct" Aggregation model and Simple Aggregation Model perform differently in the empirical study.

DEDICATION

I dedicate this dissertation to my son, Victor Ma, family, and friends who have trusted and encouraged me to complete the journey to an economist.

ACKNOWLEDGMENTS

I truly acknowledge the endless support and guidance of my advisor, Dr. Thomas A. Mroz, in completing this dissertation. I greatly thank Dr. Michael T. Maloney, Dr. Curtis J. Simon, and Dr. John T. Warner for their sincere assistance and helpful comments as members of my dissertation committee.

I am grateful to the participants of the Labor Economics Workshop for their helpful comments and support. I thank the remaining faculty members and staff in the John E. Walker Department of Economics for helping me with my through the PhD program as well as the dissertation. Finally, I thank James Von Oehsen and his team for their assistance in running my MPI jobs in Palmetto Cluster.

Grants from National Institute of Child Health and Development (R01-HD047213) provided partial support for this research.

TABLE OF CONTENTS

Page

LIST OF TABLES

List of Tables (Continued)

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Research on the estimated effect of additional resources to the local schools has direct implications for tax policies and government budgeting toward local public goods. Since it has been the focus of many policy makers in the second half of the century, it is of great importance to obtain a clear and unbiased estimation method of the value of better school resources. If school policies are undertaken without theoretical guidance, they can cause an enormous waste of society's resources. For example, the pupil-to-teacher ratio has been treated as a major force driving student outcome which leads the average pupil-to-teacher ratio in US public schools to fall from 28 students per teacher to less than 16 students per teacher over the 1940 to 1990 period. Even with this drop in the pupil-to-teacher ratio, the test scores of primary and secondary school students showed no improvement across the nation during this time period. (Hanushek, Rivkin, and Taylor 1996 [11])

In studying the effect of the school on educational reform in both the short and long run, the literature has a wide range of viewpoints and methods for estimating the school's effect on achievements. According to a recent review by Hanushek, Rivkin, and Taylor(1996) [11], there were 277 separate investigations of the school quality indicator "pupil-to-teacher ratio" and 163 studies about one of the other quality measures of schools, "expenditure per pupil". Endless research efforts notwithstanding, little consensus has been reached about the magnitude or even the direction of the school's role in a child's education. Some studies in school performance yield a simple conclusion that there is no strong or consistent relationship between school resources and student performance (Childs and Shaksshaft, 1986 [6]; Glass and Smith, 1979 [7]). Conversely, the widely publicized findings of Card and Krueger (1992) [5], together with several

other studies (Johnson and Stafford, 1973 [15], Link and Ratledge, 1975 [18], Rizzuto and Watchel, 1980 [30]) indicate that variations in school resources are related to returns to education.

To reconcile discrepancies of school effect findings, several prominent studies suggest a number of alternatives. Firstly, Todd and Wolpin (2002) [33] propose that most of the previous estimations of the child educational production function failed to accommodate the fact that educational policies and household behavior interact to determine student outcomes. The positive raw correlation between school spending and test outcome disappears when family background is controlled in the estimation. Secondly, Heckman (1996) [11] points out that endogeneity biases take effect when there are correlations among the different inputs from the school, family, and student themselves. For example, parents make systematic school choices about migration taking into account their child's education, meaning that "good" parents self-select themselves into "good" school districts. This belief confounds the question of the effect of the school doing "good" with parents doing "good" on their children's education. Omitted variable bias is worsen if the data are not rich enough to accommodate these correlations. Thirdly, Loeb and Bound (1995) [17], and Hanushek, Rivkin, and Taylor (1996) [11] show the importance of data characteristics in explaining the conflict in the findings of school effectiveness studies. They point out that significant positive school effects frequently appear in educational outcome studies using aggregate data while strong school effects are not found when micro level data are used.

This paper is aimed at addressing the various problems that exist in the literature. We specify a random parameters model for the collective decisions about school choice and education made by the family. In the joint estimation of school choice and the educational production function, we allow the educational outcome to be dependent upon the same unmeasured family preference factor that affects the family's school choice. We find that the widely used production function model with aggregated resources can cause serious specification errors that increase the endogenous bias. We conduct a Monte Carlo study comparing three endogeneity-controlling educational production models estimated under disaggregated and

aggregated school resources assumptions. We experiment with different assumptions about the heterogeneity pattern of the unobserved preference factor on the three models and find that specification errors generated by simple aggregation can be overcome by "correctly" aggregating the resources over a geographic area. In the section that follows, we empirically estimate the "Correct" Aggregation Model and the simple aggregation model using US data for the year 1994.

The paper is organized as follows: Section II provides a review of the relevant literature. Section III lists the three different modeling specifications of the joint distribution of school choice and educational production function, and section IV applies the three models to a Monte Carlo experiment. The Data used for empirical estimations are introduced in section V. Empirical estimation of the Correct Aggregation Model and simple aggregation model using 1994 data are discussed in section VI. Section VII provides concluding remarks.

CHAPTER TWO

BACKGROUND

Early studies of school effectiveness (Oates, 1969 [27]; Kain and Quigleys, 1970 [16]) focus on the linear relationship between school resources and child educational achievement or local housing value outcomes. This group of studies, which is recognized as "hedonic pricing" model or "linear test outcome projection" has a number of shortcomings. The first is related to the fact that school inputs and family inputs are both important factors affecting child academic achievement. Estimations of school effects are biased if there is no control for family background, but data on both sides for educational inputs are not frequently available. The second difficulty proven to be more crucial in the estimation is focused on the fact that school input and parental inputs are inter-correlated. The school resources a child is receiving is not "given" to him, but "chosen" by his parents considering the various family backgrounds and individual preparation issues. Hedonic pricing models fail to account for this fact because they do not deal with errors in the educational production function that are correlated with the school quality indicators. Tiebout, (1956) [31] with his sorting model, emphasizes the importance of the systematic location choice of families and the possible impact of that on the performance of the family members. Studies of housing values shows that school inputs have been capitalized into housing price (Black, 1999 [3]). Indeed, the unobserved preference factors affecting location choice of the family also affect the choices concerning their children's education, including choices among various complementary programs that aid the learning, and choices of helping the daily studies of their children. These educational efforts are usually unmeasured, and thus, not controlled for in most of the hedonic projection studies.

Hanushek, Rivkin, and Taylor (1996 [11]) derived a theoretical model to prove that the aggregation of all variables to the state level will unambiguously bias the schooling parameters

upward if these state achievement factors are omitted from the state level because the resources and state achievement factors are positively correlated. When these resources of quality indicators are ignored in the estimation, the omitted variable bias will have a serious effect on the estimation. The Coefficients of school resources are most seriously inflated when resources are aggregated to the state or a higher level. Hanushek et al. highlighted a series of reasons why the state level aggregate data behave poorly in estimating the educational production function: Schools in the United States are organized by different states (counties) and thus follow different types of policies depending on state resources/preferences. For example, it is stated that "Some 37 states have forms of teacher competency testing, while others do not, and details about the requirements for teachers vary a lot", "Policies are different across states for teacher tenures too", "States also vary in terms of requirement for graduating with high school diploma" (Hanushek, Rivkin, and Taylor (1996 [11]). While controlling for the unmeasured factors is difficult, these unmeasured state regulatory factors bias the estimates if not controlled for in the estimation.

Loeb and Bound (1995) [17] found large school effects using division level aggregation data. They argue that the difference in data characteristics could crucially affect estimation "more than the differences in outcome measures or biases from labor market influences" (Loeb and Bound (1995) [17]). They follow most studies to claim that aggregation decreases the endogeneity problem of parents selecting the school attendance area when the selection issues are not controlled. Their second argument in support of aggregation models is about the measurement error. It is well-known that empirical studies using survey data suffer from measurement error, due to the fact that the local respondent might not know the meaning of each questionnaire or that the value that has been asked for cannot be measured precisely. One way of eliminating the inconsistency of estimators caused by the measurement error is using Instrumental Variables Since it is correlated with the true value of the regressor but not correlated with the random measurement error, aggregating within groups of resources is like creating an instrumental variable. Previous studies find that group averages reduces the bias if it raises the signal-to-noise

ratio.(Hanushek, Rivkin, and Taylor (1996 [11]). Even though parents are more likely to look at a school district when making the location choice decision (since a house located within the boundary of better school districts can be much more expensive than houses within the same area but outside the boundary), they believe that the value added approach used by most micro data studies can help more than the aggregated data to decrease the endogeneity problem.

Our study tries to reconcile the previous contradictory findings of schools effect on child academic achievement. We contribute to this area of literature in multiple ways. By modeling the location/school choice of parents, our analysis framework has certain advantages over models that have do not control over school selection. Firstly, it emphasizes the importance of the endogeneity of school qualities on the estimation of the effect of schools on student academic achievement. Since parents valuing better schools raise the other (unmeasured) parental inputs to their child's education, the unmeasured family preference on school choice affects the student test outcome. Secondly, it has the ability to account for the local/state level unobserved school quality indicators since families look at both the observed (pupil-to-teacher ratio, teacher salary, dropout rate, etc.) and unobserved (state regulation on teacher competency test, etc.) factors when choosing a school district for their children. Thirdly, with this endogeneity issue controlled for in the model, we access the performance of different modeling specifications (aggregated versus disaggregated) on the estimation of child educational production function. We explore a plausible modeling specification that "correctly" aggregates school resources over particular geographic areas. This model successfully deals with two aspects of aggregation bias existing in the prominent examples of school effective studies: 1) the bias caused by the aggregated local public goods that a family is comparing from the choice set when they make their residential decision. 2) the bias caused by the aggregated school resources that have been accounted for in the child production function by linking to an individual, and then estimated by controlling the endogeneity of school choice of a family.

CHAPTER THREE

MODELING SPECIFICATIONS

In this chapter, we model the school choice behavior of families. Based on the quality of the chosen school district, the student's educational production function is modeled. The aim is to control the endogeneity of school quality in the child educational production function and correctly aggregate the district information if aggregation over a certain geographic region is required.

3.1 Full Location Information Model

Families are assumed to make their location choice according to a school district's boundary. Each family with a child over the age of five is faced with choosing from all school districts (J) in the United States. Since about 34% of US counties contain more than one school district, consider the following model. Suppose that the nation has k counties where k = {1,…,K}. For each k labeled county, there are J_k school districts. We follow a timeline to mimic the real world by assuming that families make a school choice.

The test score outcome for the children is achieved conditional on the chosen school district's qualities. Families make a location choice within a typical school district boundary. In particular, they consider the school quality of the destination Ω_{j_k}, the expected income from the location, which is idiosyncratic to the parents characters, and the travel distance between the current location of the family and the possible location from the choice set Z_{i,j_k}. An unobserved preference factor, μ_i, that the family has toward the school qualities also affects the decision. The expected utility of choosing each location is specified as:

$$U_{i,j_k} = U_{i,j_k}(\Omega_{j_k}, Z_{i,j_k}, \mu_i) + \varepsilon_{i,j_k} \tag{3-1}$$

Assuming the error term ε_{i,j_k} follows an i.i.d. extreme value distribution, the probability of this family *i* choosing school district j_k for their children, after conditioning on the unobserved preference factor μ_i, is specified as a standard conditional logit formula:

$$\Pr(j_k = j_k^* \mid Z_i, \Omega_{i,j_k^*}, \mu_i) = \frac{EXP[U_{i,j_k^*}(\Omega_{j_k^*}, Z_{i,j_k}, \mu_i)]}{\sum_{j_k=1}^{J} EXP[U_{i,j_k}(\Omega_{j_k}, Z_{i,j_k}, \mu_i)]'} \quad . \tag{3-2}$$

The choice of location determines the school inputs for the family. Student outcome is a function of the personal specific characteristics and student preparations (X_i) such as the age and race of the child, the mother's AFQT score, the mother's working status, characteristics of the quality of school his/her family chooses for him (Ω_{i,j_k^*}), and the same family specific preference factor μ_i that affects the utility function (3-1):

$$test_i = f(\Omega_{i,j_k^*}, X_i, \mu_i) + v_i \tag{3-3}$$

Assuming the error term v_i follows an i.i.d. normal distribution, the likelihood of observing a test score $test_i^*$ achieved can be specified as the standard normal distribution formula conditional on the unobserved heterogeneity factor μ_i :

$$f(v_i) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left[\frac{(test_i - f(\Omega_{i,j_k^*}, X_i, \mu_i))^2}{2\sigma_v^2}\right] \tag{3-4}$$

From simple probability theory, the joint distribution of achieving student outcome and choosing the location can be written as the product of equations (3-2) and (3-4) if the two vector of error terms ε_{i,j_k} and μ_i are independent. However, as stated in the last section, school characteristics Ω_{i,j_k^*} is endogenous in the above child educational production function, the family's chosen school district, and the same unmeasured preference factors that affect the

location choice of the family could also affect the student test outcome. For example, we could imagine parents who really care about the academic performance of the children residing in their best affordable school district, and at the same time, spend as much time as possible helping their children with their homework to improve the student test score outcome. Heckman (1996) [13] points out that most studies in this area suffer from omitted variable bias because they fail to control possible unmeasured variables that may affect both test outcome and location choice.

To control for the unobserved factor μ_i that may cause omitted variable bias found in studies similar to this one, we use a random parameters specification for the utility function determining location choice. Together with having this preference factor in the utility function, we specify a linear contribution of this unobserved factor toward the student's test score. Our functional specification is very similar to McFadden and Train (2000) [30]'s mixed logit random parameters specifications that allows the preference of the family on school characteristics to vary across individuals. On top of that, we estimate the heterogeneity points with their possibilities to be observed. Therefore, the marginal contributions of schools to student outcome are allowed to be differentiated across families as well.

To elaborate on the random parameter specification used in this model, suppose that the utility function determining the best school choice of parents is linear in location specific characteristics idiosyncratic to each family, and school quality indicators Ω_{i,j_k}:

$$V_{i,j_k} = \alpha_1 Z_{i,k} + \alpha_2(\mu_i)\Omega_{i,j_k} + \varepsilon_{i,j_k} \tag{3-5}$$

Where $\alpha_2(\mu_i) = \alpha_2^0 + \alpha_2^1 \mu_i$,

ε_{i,j_k} follows i.i.d. extreme value distribution. In this specification, parents are differentiated in the importance they place on the school inputs in their choice of a place of residence.

In the child educational production function, test score is assumed to be linearly dependent on the chosen school characteristics, while the family preference factor μ_i also contributes to the test outcome:

$$test_i = (test_i^* \mid \Omega_{i,j_k^*}) = \beta_0 + \beta_1 X_i + \beta_2 \Omega_{i,j_k^*} + \rho\mu_i + \nu_i \quad (3\text{-}6)$$

where ν_i is assumed to be i.i.d. normally distributed. The parental preference over school characteristics could also be an unobserved input to the child educational production function.

After controlling for the common factor μ_i in utility and production function, the two error terms ν_i from (3-5) and $\varepsilon_{j_k,i}$ from (3-5) will be mutually independent. Conditional on this common factor μ_i, the unconditional likelihood function can be obtained by integrating over the distribution of the common factor. Assuming the distribution of the unobserved heterogeneity factor is discrete, with finite supporting points for the distribution of μ_i, together with estimating the probability of each point of support, we discretely approximate the true likelihood function by specifying (Mroz 1999 [24]):

$$L_i = \sum_{h=1}^{H} \Pr(\mu_{h,i}) \{ \Pr(j_{k,i} = j_{k,i}^* \mid \Omega_{i,j_k}, Z_{i,j_k}, \mu_{h,i}) f(test_i = test_i^* \mid \Omega_{i,j_k^*}, X_i, \mu_{h,i}) \} \quad (3\text{-}7)$$

Assuming the exact school choice of the family is known, this is what we call Full Location Information Model.

3.2 Correct Aggregation Model

If the data contain the exact school district chosen, then the fully observed location model can be implemented empirically. However, due to data limitations on school studies like this one, economists frequently aggregate school resource data into some higher level. For example, because of the confidentiality concern from the NLSY national survey, we only know which county the surveyed family is located. However, we could obtain school district level quality

information from the Common Core Dataset (CCD). Many studies use the county level or state level average "school district qualities" to represent the quality of school the child is attending. This simple aggregation has been proven to be ineffective with exaggerated omitted variable bias (Hanushek, Rivkin, and Taylor 1996 [11]). We provide a solution to this problem.

There are two differences of the aggregation model from the fully observed location model. First, only the county of residence is known to the researcher, not the exact school district; second, while we do observe the test score for each child, we do not know the exact school characteristics the child receives. To follow this reality, we assume the family considers the optimal school district across all counties in order to make a location choice so that the optimal county choice is the outcome of the sum of the probabilities of choosing each school district within the county. In the fully observed location model we know precisely in which school district someone lives. Denote this district as j_k^* in county k. Then, the probability of observing the chosen district and test score in that district can be written out to be:

$$\Pr(j_k^* = j_{k,i} \,\&\, test_{i,j_k*} = test^*_{i,j_k*}) = \Pr(j_{k,i} = j_k^*) * f(test_{j_k*,i} = test^*_{i,j_k*}) \qquad (3\text{-}8)$$

All that is known here is that county k was chosen and the child's test score of that county. To obtain the probability of observing a chosen county of residence and a child's test score outcome, we integrate $\Pr(j_{k,i} = j_k^*) * f(test_i = test_i^*)$ over all of the school districts in county k:

$$\Pr(k_i = k_i^* \,\&\, test_{k,i} = test_{k,i}^*) = \sum_{j_k=1}^{J_k} [\Pr(j_{k,i} = j_k^*) * f(test_i = test_i^*)] \qquad (3\text{-}9)$$

We can look at this in specific linear functional form as well. Assuming the chosen school district J_k^* is known to the researcher, the probability that person i chooses school districts J_k^* can be specified as:

$$\Pr(j_{k,i} = j_k^* \mid \Omega_{j_k*}, \mu_i) = \frac{\exp\left[U_{i,j_k*}(\Omega_{j_k*}, Z_{i,j_k}, \mu_i)\right]}{\sum_{k=1}^{K}\sum_{j_k'=1}^{J_k}\exp\left[U_{i,j_k'}(\Omega_{j_k'}, Z_{i,j_k}, \mu_i)\right]'} \tag{3-10}$$

with linear relationship between inputs and utility:

$$\Pr(j_{ki} = j_k^* \mid \Omega_{j_k}, \mu_i) = \frac{\exp\left[\alpha_1 Z_{i,j_k'} + \alpha_2(\mu_i)\Omega_{i,j_k*}\right]}{\sum_{k=1}^{K}\sum_{j_k'=1}^{J_k}\exp\left[\alpha_1 Z_{i,j_k'} + \alpha_2(\mu_i)\Omega_{i,j_k}\right]'} \tag{3-11}$$

If the county choice is known to the researcher, denoted as k_i^*, we sum the probability of choosing each school district within the county ($j_{k^*}, j = 1,.....J_{k^*}$) to obtain the probability of observation i to choose that county and the test score outcome to be linked to each school district in the county:

$$\Pr(k_i = k_i^*) * f(test_i = test_i^*) =$$
$$\sum_{j_{k^*}=1}^{J_{k^*}}\left[\frac{\exp\left[\alpha_1 Z_{i,j_k'} + \alpha_2(\mu_i)\Omega_{i,j_{k^*}}\right]}{\sum_{k=1}^{K}\sum_{j_k'=1}^{J_k}\exp\left[\alpha_1 Z_{i,j_k'} + \alpha_2(\mu_i)\Omega_{i,j_k}\right]'} * \frac{1}{\sqrt{2\pi}\sigma_v}\exp\{\frac{[test_i - (\beta_0 + \beta_1 X_i + \beta_2\Omega_{i,j_{k^*}} + \rho\mu_i)]^2}{2\sigma_v^2}\}\right] \tag{3-12}$$

Note that in this specification, school quality information is available for each of the school district level even though it is not necessarily linked to a student in that district. Under discrete heterogeneity assumptions about the unobserved preference factor μ_i, if estimating both the possibility of choosing the county of residence and achieving the test score conditional on choosing the county we model, then the likelihood function is given by:

$$L_i = \sum_{h=1}^{H} \Pr(\mu_{h,i}) \{ \sum_{j_{k^*}=1}^{J_{k^*}} \Pr(j_{k^*,i} = j_{k^*,i} \mid \Omega_{j_{k^*},i}, Z_{i,j_k}, \mu_{h,i}) \Pr(test_i = test_i^* \mid \Omega_{j_{k^*},i}, X_i, \mu_{h,i}) \} \tag{3-13}$$

Note that this model is different from the fully observed location model by a summation of joint probabilities within each county. Even though only one test score is observed and location choice is known at the county level, this model maintains as much information as possible by using school district level quality information and "correctly" specifies a family's incentive to move to a particular location. This property of the model is consistent with the individually observed Full Location Information Model presented in the last section. It should yield estimations with less bias than the commonly used simple aggregation model discussed in the last section. Thus, we call this model the "correct" aggregation model.

3.3 Ad Hoc Simple Aggregation Model

When location choice is partially observable to the researcher, the simple aggregation model conveniently assumes that a family makes their county choice through considering the county level average school qualities. The child educational production function inputs will be simply aggregated to the county level. It is a widely used method in the literature and has many supporters from all aspects of applied economic studies. Now, we look at its application in the school choice and educational production function in detail. The purpose of this discussion is to mimic the use of aggregate inputs by many researchers in practice.

Assume there are K counties in the nation, and for any county k, there are J_k school districts in the county: $j_k = 1.....J_k$, $k = 1.....K$. If we have the information on the pupil-to-teacher ratio, expenditure per pupil, and teacher salary per pupil for each school district j_k in county k, we assume the utility from choosing county k is only a function of the county level average school characteristics:

$$\Omega_{j_k} = \Omega(\frac{pupil}{teacher}_{j_k}, \frac{salary}{pupil}_{j_k}, dropout_{j_k}), \tag{3-14}$$

Following previous studies using group average resources as the resource received by the individual, we take the average of these ratios for school districts within each county and obtain the county average school quality indicators:

$$\overline{\Omega_k} = \Omega(\overline{\frac{pupil}{teacher}}_k, \overline{\frac{salary}{pupil}}_k, \overline{dropout_k}) \tag{3-15}$$

Assuming the errors are i.i.d. extreme value distributed, the probability that person i choosing county k is a function of the average school characteristics in county k and other factors idiosyncratic to the person and location (dependent on the preference factor) that might affect the location choice of the family is:

$$\Pr(k_i = k^* \mid \overline{\Omega_k}, \mu_i) = \frac{\exp\left[\alpha_1 Z_{i,k^*} + \alpha_2(\mu_i)\overline{\Omega_{i,k^*}}\right]}{\sum_{k'=1}^{K} \exp\left[\alpha_1 Z_{i,k'} + \alpha_2(\mu_i)\overline{\Omega_{i,k'}}\right]'} \tag{3-16}$$

Note that the aggregation causes some information loss for the production function through the loss of the variations in Ω_{i,k^*}.Test score outcome, under the aggregation assumptions, is based on the average school quality of the chosen county $\overline{\Omega_k}$ along with other background factors (X_i) of the student:

$$test_i = \beta_0 + \beta_1 X_i + \beta_2 \overline{\Omega}_{k^*,i} + \rho\mu_i + \nu_i \tag{3-17}$$

The likelihood functions for the aggregation case can be written by specifying a discrete approximation of the distribution of the unobserved heterogeneity μ_i :

$$L_i = \sum_{h=1}^{H} \Pr(\mu_h)\{\Pr(k_i = k_i^* \mid \overline{\Omega}_{k,i}, Z_{i,k}, \mu_{h,i}) \Pr(test_i = test_i^* \mid \overline{\Omega}_{k^*,i}, Z_{i,k}, \mu_{h,i})\} \tag{3-18}$$

Unlike the Correct Aggregation Model, this model cannot be derived from the Fully Observed Location Model. If we only observe the chosen county of residence (or city, state,

division), parents should look at a specific school district (ideally the best one maximizing the utility function) within a county when the family is making a migration choice instead of the county average school quality as the main concern when relocating. Also, the student test outcome should be linked to the school resources they receive instead of county average school qualities. About 34% of the US counties have one school district, but more than 65% of US counties have two or more school districts. In some counties we observe more than ninety school districts with differentiated structures of management and educational attainment. Therefore, we believe the Correct Aggregation Model should provide more accurate estimates of school effects than the Ad Hoc Simple Aggregation Model because it correctly accounts for the fact that families look at the best school district quality in a county when choosing a county to reside. In the following Monte Carlo study that compares the two models, we create a world that is similar to the choice set of unified school districts in the United States. We have 2000 children in the sample choosing from a subset of counties that mimics the real US counties to reside, and for each county, there are various numbers of school districts.

CHAPTER FOUR

DATA

This chapter details the data preparation for the empirical estimation of the Correct Aggregation Model and the Simple Aggregation Model. Because the exact school district choice but not the county choice of the family is reported in the NLSY national survey, we have only these two models under aggregation bases from which to choose. There are three sets of data that we obtained from different resources to form the structural model. We obtain 1) the school district character information used in both utility function and production function estimation from NCES CCD data files; 2) the expected wage offer for the mother and the expected earnings offer for the father from Census/CPS, and 3) the child and family background data from NLSY_Youth, NLSY_Child Supplement. Details about data files we use are listed below:

4.1 School District Resource Data: Common Core Data Files from NCES

Geographic identification of the residential community and thus the school district a child is attending is very important when estimating the educational production function as part of a structural model considering location choice. School district characteristics considered to be the major factors affecting the utility function of families are Pupil-to-teacher ratio, Expenditure per Pupil and Dropout Rate of the local school district. We obtain the national school district data from the National Center for Education Statistics (NCES) Common Core of Data (CCD) 2000. The Common Core data files (CCD) from the NCES provide more than 20 years of school level, school district level and state level local education agency information. It has three major survey categories for the local school district level data: Local Education Agency (School District) Universe Survey Data: 1986–Present; Local Education Agency (School District) Finance Survey (F-33) Data: 1990–Present; Local Education Agency (School District) Universe Survey Dropout and Completion Data (1991–Present). For every local school district agency, there is a consistent

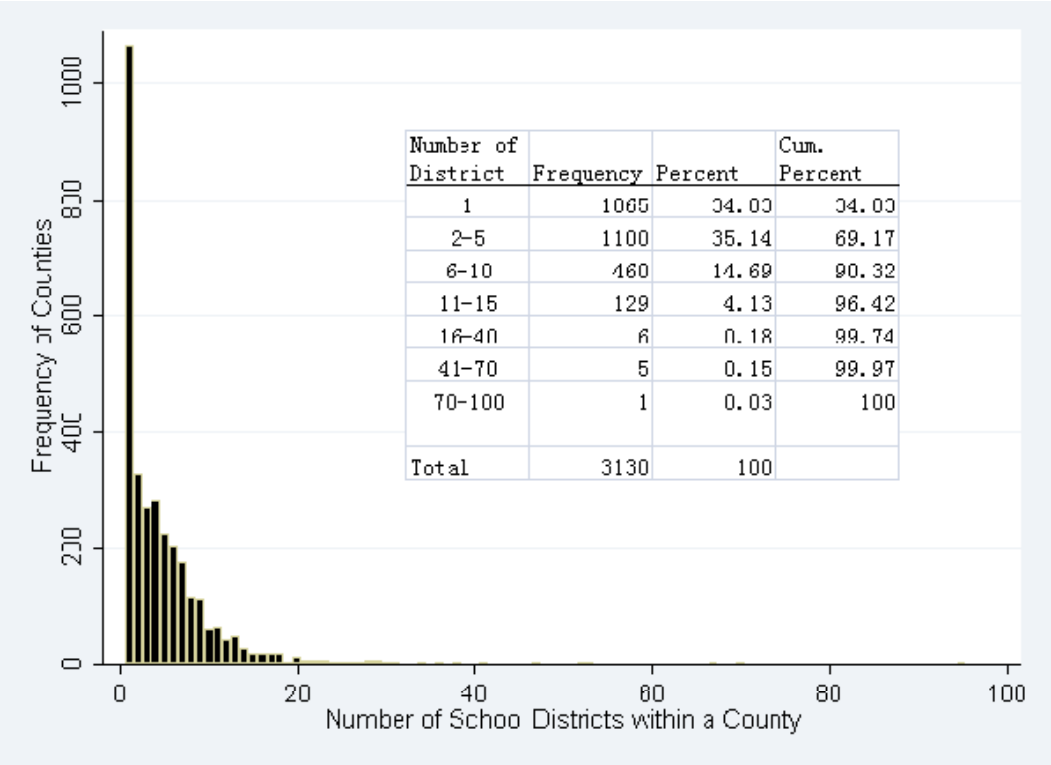
ID code called LEAID (Local Education Agency ID) assigned by NCES to the agency. This dataset is used extensively for researches concerned with local educational achievement, budget, expenditures, local housing market and local labor market supply issues in the United States.

Our variables of interest are the number of students, number of teachers, total expenditure and total dropout number with total dropout base for each school district. Those variables are then used to construct the ratios we utilize, which are pupil to teacher ratio, expenditure per pupil and dropout rate. To define the local education agencies that are distinguished from the administrative agencies or vocational services that are for the adults and having degree offers up to 12, we get to know the geographic location of the school districts. To do this, we use the 1990 Census School District Map. The Map has 15,512 local education agencies where 12920 of the 15,512 are defined as the "Unified" or primary/secondary school districts that offer a degree up to high school diploma. "Standalone" school districts account for 3,582 of the total school districts where only partial degrees (not up to 12) are offered. For each of these 3,582 standalones in the map, it is easy to find a nearby unified school district where a certain student might go attend within the county by latitude and longitude geographic information. Finally, we have a data file identifying the up-to-12th-grade school districts in the United States with their area of occupation, district boundaries, degrees offered, and other important quality indicators defined and identified. A detailed description of how we define the US school districts across to 1987-2005 time period and deal with the problem of missing data in the original survey is included in *Appendix C.*

Figure 4-1 shows the distribution of school districts among US counties in 1994. According to the Census School District Map, 1,065 counties have only one school district which accounts for approximately 34% of the county file. We also have 1,100 counties that have two to five school districts per county, and about 600 hundred of them having six to 16 school districts per county. Many counties have a number of school districts with very different quality levels within the county. For example, "Cook County" in Illinois has 95 school districts where the pupil-to-teacher ratio of these school districts ranges from 8 students per teacher to 21 students

per teacher; “Bergen County” in New Jersey has 70 school districts, and the expenditure among these school districts goes from 1,400 dollars to 9,000 dollars per pupil. *Table 5-1* provides the summary statistics for the local public school dataset that we use. Generally, class size has a mean of 16 students in 1994, expenditures per pupil averages at 4206 dollars per student, and overall dropout rate over the first to 12^{th} grade ranges from zero to 30%, averages at 2%. We show the division of within and between county standard deviation of all school districts as well. The small difference in the standard deviations of all quality indicators shows that the school district qualities are not unanimous within a county. This statistic indicates that the Correct Aggregation Model can yield different estimation result from the Ad Hoc Model because of the unbalanced distribution of school district across counties.

Figure 5-1 Distribution of School Districts among US Counties in 1990



Source: 1990 Census, NCES CCD Universe Survey 1990

Table 4-1 Summary Statistics for Local School Districts for 1990

Variable		Mean	Std.	Observations
Pupil Teacher Ratio	Overall	15.661	3.769	N = 14765
	Between Counties		2.853	n = 3130
	Within Counties		2.125	T-bar = 4.71725
Expenditure per Pupil($1000)	Overall	4.206	1.873	N = 14765
	Between Counties		1.266	n = 3130
	Within Counties		1.371	T-bar = 4.71725
Dropout Rate	Overall	0.022	0.022	N = 14765
	Between Counties		0.020	n = 3130
	Within Counties		0.014	T-bar = 4.71725

Source: NCES CCD Universe Survey, Finance Survey, Dropout-Completion Survey for Local Education Agency in year 1990

4.2 Labor Market Identification for Expected Wage/Earnings Rate for the Parents

When parents make decisions about where to live, they consider local environmental issues such as the expected income and job opportunities together with the local public goods such as the school qualities. We follow the Census/CPS definition of labor markets that offer expected income to the parents and made some arrangement of the data to better conform to the reality. Details about our definition of labor markets are discussed in *Appendix E.1*. Generally, the Census Bureau has a variable "PUMA" which shows household's county group of enumeration. "PUMA" are contiguous areas with a combined population of 100,000+ residents after the 1980. We use METAREA as the relevant labor market for counties that are "urban." (in a METAREA). We use the observation's (CONS) PUMA code plus the STATEFIP code to identify the local labor market for counties that are non-urban. We obtain in total 496 labor markets in the US: 272 are urban, and 224 are rural. For the 224 rural areas, we use CONSPUMA to define a labor market within a typical state.

For each labor market we define, we use local linear regressions to obtain an expected wage/earnings offer for every combination of sex, age, education, and race. Details about the definition of the labor market and the Local Linear Projection method we use for expected wage rate for the NLSY parents can be found in *Appendix E.2*. We have extracted a mapping file from the geographic files of IPUMS that link each county in the US 1990 to its respective labor market. Each male observation in our sample from NLSY has an expected annual earnings rate at any possible location he might choose to locate the next biannual time period. The same is true for female wages at each possible location. In each potential labor market, we obtain one expected wage for part-time female workers and one for full-time female workers. Usually, we have a labor market that includes several school districts because it is largely possible for a family living in several school districts around a labor market to go to the center metropolitan area to work. It is reasonable for the parents living in separate counties to share a common labor market, but their

wage/earnings offer are not unique in this labor market, instead they follow a distribution depending on their own age, sex, educational background, and race combinations.

Table 4-2 describes the summary statistics of wage/earnings offer for NLSY parents. This is obtained by merging the NLSY parents' demographic characteristics with the wage/earnings prediction file we have for year 1994. From the table we could see that the age of mother of the young child ranges from 30 to 37, age of father ranges from 22 to 59. The grade of the fathers is roughly at the same level of the mothers. Mothers do not see a huge hourly wage difference if they choose to work on a full-time or part-time basis. Full-time annual earnings for fathers are averaged at roughly $25,500 across the locations.

Table 4-2 Summary Statistics of NLSY Mother and Father Wage/Earnings Offer from Locations

In Year 1994

Variable	Mean	Std. Dev.	Min	Max
	Mother's Wage Offer from Locations			
Age	33.500	2.291	30	37
Nonwhite	0.500	0.500	0	1
Grade	12.667	2.494	10	16
Ln Part-time Wage($/hr)	2.255	0.313	1.461	3.111
Ln Full-time Wage($/hr)	2.347	0.302	1.779	3.178
	Father's Earnings Offer from Locations			
Father Age	38.558	8.765	22	59
nonwht	0.494	0.500	0	1
Grade	12.532	2.314	10	16
ln Full-time Annual Earnings ($/yr)	10.146	0.487	7.444	12.305

Wage and Earnings are in 2005 dollars.

4.3 Child and Family Data Source

We use the Geocode version of NLSY79 (the National Longitude Survey of Young Adults) dataset and its Child-Mother supplement as main data files for the student's background and test score information in the child educational production function. The NLSY79 began in 1979 with a national sample of 12,686 young adults between the age of 14 and 21. Beginning in 1986, the NLSY-Child collected data on all of the children born to the female NLSY respondents. The 1994 NLSY combined sample used here supplies data on children less than 15 years old with mother's age between 31 and 37 at the end of 1994. Merging the main Youth file extracted from the NLSY79 main Youth dataset and the child-based file extracted from the NLSY79 Child&Young Adult Data, we get a coherent dataset that has longitudinal information of a child on the aspects of maternal inputs, family background, educational preparation, etc. because the interview for both child and mother is conducted biennially, each period in our empirical model corresponds to a 2-year interval. Here we use the family background and test score information for year 1994 (1992 test score for previous period math score) to conduct the empirical study.

We use student PIAT math score as the indicator of the children's educational achievement. We use the value added approach by controlling for the student's last period PIAT math score. Out of the 3618 children we have in 1994, 3576 children in the year 1994 reports their math score, while 2584 of reports their math score in year 1992. The math score used here are the tentile score from 1 to 10, not the exact score. We link every test score in the sample to a school district according to the county of resident information they provide in the survey. Other than school quality factors such as pupil teacher ratio, expenditure per pupil and the dropout rate of the school district these children studies, there are family side student background factors directly affecting the student test outcome. We have also include mother's AFQT score, race, an indicator variable for boy, age of the student, whether mother works full-time or part-time to the educational production function.

Another factor directly affects both the production function and the utility function is whether the father is present in the family. In choosing a location to reside, the expected earnings for fathers are not taken into consideration if the mother is the head of the household. We don't consider this difference in the Monte Carlo study in *chapter 5*, but we arranged a change of functional form to accommodate the single mother families in the empirical study in *chapter 6*. More importantly, we argue that whether a father is present in the family takes an important effect on the children's educational achievement. The NLSY survey of Youth provides answers to questionnaires regarding the surveyed woman's family members. The information about the woman's family members (up to 15) is stored in the part 'Household Record'. Specifically, it provides each household member's: Relationship Code to Youth, sex, age, grade and working or not information. The Relationship code shows whether this is a "spouse" or "partner" or any other possible relationship of this family member to the surveyed youth. We take both the "spouses" and the "partners" as the indicator for the presence of the father. We also use the father's age and grade information to find the expected earnings information for the father. Out of the 2584 children who provide PIAT math score in the year 1994 database, there are 1731 children that had a father. *Table 4-3* gives the summary statistics for the Child and Family Information dataset we use.

Table 4-3 Summary Statistics for Child and Family Data

Variable	Observations	Mean	Std. Dev.	Min	Max
Age of Child	2584	10.625	2.180	7	15
Boy	2584	0.497	0.500	0	1
AFQT Score of Mother	2584	0.456	0.273	0	1
Nonwht. Child	2584	0.559	0.497	0	1
Mother Full Time Work	2584	0.438	0.496	0	1
Mother Part Time Work	2584	0.309	0.462	0	1
Age of Mother	2584	33.546	2.167	30	37
Mother High School Grad.	2584	0.495	0.500	0	1
Mother College Grad.	2584	0.331	0.471	0	1
Father Age	1731	35.837	4.865	22	59
Father Grade	1731	12.287	1.544	10	16
PIAT Math Score	2584	5.372	2.858	1	10
Previous Period PIAT Math	2584	5.420	2.840	1	10

The constricted version of NLSY79 gives each interviewed family's county of residence, but not the school district that the child is attending. Unfortunately, in almost all the studies of US nationwide school resource effects, specific resources cannot be directly matched with the students receiving them due to the lacking of a coherent dataset that gives all family background and school resources. This frequently implies that aggregation in the dataset when merged from different resources, and that is the initial incentive of us raising this question about difference in estimation when using aggregated or disaggregated data. In this study, we do not observe migration within the county. Distance information between family's current location and prospect location is obtained by latitude/longitudinal great circle distance between the centers of two counties. In the estimation of Correct Aggregation Model, the possibility of choosing a county to reside is the sum of possibilities choosing each school district in that county. In the next section, we conduct a Monte Carlo study for the Full Location Information model, Correct Aggregation Model and the simple ad hoc aggregation model using data collected from resources described above. We demonstrate that the Correct Aggregation Model is superior to the simple ad hoc aggregation model, in terms of both standard errors and the mean square errors. in the session that follows, we compare the two aggregation models' empirical performance. The results are discussed in the final section.

CHAPTER FIVE

THE MONTE CARLO

In this section, we conduct a Monte Carlo study using all three models described in the *Chapter 3* and compare their performance with different assumptions about the unobserved preference factor of the parents toward school districts. By replicating each data generation-estimation process 50 times, we assess the performance of all three models separately under each of the following assumptions: 1) no variation in the unobserved preference factor for the family, 2) five discrete points supporting the unobserved preference factor, and 3) continuous random normal distribution of the unobserved preference factor. We use the actual real pupil teacher ratio, expenditure per pupil, and dropout rate for school district quality information obtained from NCES 1990 in Ω_{j_k} in the fully observed location model and the Correct Aggregation Model. We aggregate the data to the county level and use county average pupil-to-teacher ratio, expenditure per pupil, and dropout rate in $\overline{\Omega_k}$ in the Ad Hoc Simple Aggregation Model.

We also evaluate these models' performance under different sub sampling drawing criteria of the locations to assess McFadden (1976) [19]'s postulation that with conditional/multinomial logit framework, estimation can be performed on a subset of alternatives without inducing inconsistency. If using all US counties as the choice set, we have more than 3,000 counties with more than 15,000 school districts in this country. The random sampling procedure proposed by McFadden states that one could create a subsample set that contains the chosen alternative together with a randomly drawn subset of alternatives from the original choice set of counties with school districts. Suppose in the choice set there are K counties with a total of J school district (a school district in county k is denoted $j_k = 1,....J_k$), and C is a subset of K ($C \subseteq K$) (where a school district in county c is denoted $j_c = 1,....J_C$). Assume the optimal

school choice to be j_k^* ($j_k^* \in J_C$), then the probability of making the school choice j_k^* from the full choice set K can be described as the probability of choosing j_k^* from the subset C:

$$\Pr(j_{k,i} = j_k^* \mid \Omega_{j_k*}, \mu_i) = \frac{\exp\left[U_{i,j_k*}(\Omega_{j_k*}, Z_{i,j_k}, \mu_i)\right]}{\sum_{c=1}^{C}\sum_{j_c'=1}^{J_c}\exp\left[U_{i,j_c'}(\Omega_{j_c'}, Z_{i,j_c}, \mu_i)\right]'} \tag{5-1}$$

If using the county average school qualities in writing the probability of choosing a county, then the function (5-1) under sub sampling assumptions should be rewritten as:

$$\Pr(k_i = k_i^* \mid \overline{\Omega_{k*}}, \mu_i) = \frac{\exp\left[U_{i,k*}(\overline{\Omega_{k*}}, Z_{i,k}, \mu_i)\right]}{\sum_{c'=1}^{C}\exp\left[U_{i,c'}(\overline{\Omega_{c'}}, Z_{i,c}, \mu_i)\right]'} \tag{5-2}$$

Note that in the logit estimation underlying the random sub sampling procedure, the denominator has only the exponential of linear expressions of characteristics of the subset of alternatives. The purpose of the Sub Sampling technique is to relieve the burden of calculating the utilities from all expected locations in the choice set. Consistency of the resulting maximum likelihood estimators relies on the Independence of Irrelevant Alternatives (IIA) property of the error terms in the discrete choice model. In this Monte Carlo study, we assess the IIA property of the log likelihood model of choosing school districts and achieving a test outcome by using sub sample county numbers of 10, 15, 25, 50, and 100 counties from the full choice set of 276 counties. We show that the sensitivity of the production function estimates and the utility function estimates under the different sub sampling assumptions and compare the results in this chapter.

5.1 Data Generating Process in Monte Carlo Experiment

To make the comparison of aggregated and disaggregated models more obvious and to simplify the data generating in the Monte Carlo experiment, we assume all the factors concerning location choice and test score outcomes are the school qualities. Interactions with the individual

characteristics are of course helpful in the empirical implementation but should not cause significant result reversals in the Monte Carlo experiment. For the location choice, the utility function is simulated by the sum of 1) a known part that is a linear function of the above three characteristics, 2) a heterogeneity part (which represents the unobserved family preference factor for school inputs) that interacts with the school characteristics, and 3) error term that distributed i.i.d. extreme value in all choices. More specifically, we define equation (3-5) in Chapter 3 to be:

$$U_{i,j_k} = -0.5(1+\mu_i)\frac{pupil}{teacher}_{j_k} + (1+\mu_i)\frac{Expenditure}{pupil}_{j_k} - 0.5(1+\mu_i)dropout_{j_k} \quad (5\text{-}2)$$

Most previous studies (Loeb and Bound (1995) [17], and Hanushek, Rivkin, and Taylor (1996) [11]) found the marginal contribution of the three quality factors of schools are in the range of zero to 0.50, so a coefficient scale of -0.5 to 1 is used as "true" coefficients in this study. The outcome of choosing the school district is generated by defining the maximum value emerging from a sorting through the utilities that each location gives according to equation (5-2).

The generation of test score outcome is based on 1) the chosen (optimal) school district characteristics together with 2) the heterogeneity preference component, and 3) an i.i.d. standard normal random error term.

$$test_i = (test_i^* \mid \Omega_{j^*,i}, j^*) = 1 + 0.5*(-\frac{pupil}{teacher}_{j_k} + 2*\frac{Expenditure}{pupil}_{j_k} - dropout_{j_k}) + 1.0*\mu_i + \nu_i \quad (5\text{-}3)$$

We use the real data on local educational agencies to conduct this Monte Carlo Study. The original data is drawn from the NCES Common Core Survey of School Districts in 1990. Detailed description of the dataset is listed in Chapter 4. In the original data files, we have 11,392 school districts in 1990, across roughly all 3,130 counties in the United States. There are various numbers of school districts in each county with about 34% of counties having only one school district. For the purpose of examining the two aggregate models' ability to recover the true coefficients, we randomly draw 276 counties with a total of 1,494 school districts from the

original 3,130 counties file to form the choice set of school districts in the data generating process. We have the drawing of counties (with school districts) subject to the constraint of no more than 10 school districts within a county, but maintain a mean of 5 school districts across the counties. We keep the choice set of counties that have averaged 5 school districts for the Monte Carlo to clearly distinguish the Full Location Information Model, the Correct Aggregation Model, and the Ad Hoc Simple Aggregation Model. From the description of models in the last section, the three models are only differentiated from each other for counties having more than one school district, due to the fact that county average resources are the same as individual school districts resources if only one school district within a county exists. We adjust the scale of Dropout Rate (in percentage) and Expenditure per Pupil (in thousand dollars) to keep a roughly matching scale of the parameters. We also drop observations three standard deviation away from the mean to keep a reasonable dataset. This also causes some counties with one school district to drop out from the choice set. We obtain an average of 5.5 school districts per county in the final choice set.

Part 1 of *Table 5-1* shows the distribution of school districts within counties. Part 2 of the table shows the summary statistics for the three major indicators of school quality. The average expenditure per pupil is roughly $6000 for the year 1990, we have a mean of 15 students per teacher, and the dropout rate ranges from 0.11% annually to 13.99% annually overall. Standard deviations of the three school quality indicators within and between counties are also shown in *Table 5-1*. Notice that for pupil-to-teacher ratio and expenditure per pupil, variance between and within counties are relatively similar. This shows that school districts within a certain county are different in terms of structure of class size and the financial situation. The variable Dropout Rate has a variance value much smaller within counties than between counties, which means that this variable is relatively constant for school districts within a certain county.

Table 5-1 Summary Statistics for School District Data used in Monte Carlo

Part 1: Frequency of school districts

Number of District	Frequency	Percent	Cum. Percent
1	29	10.51	10.51
2	32	11.59	22.1
3	25	9.06	31.16
4	35	12.68	43.84
5	27	9.78	53.62
6	19	6.88	60.51
7	29	10.51	71.01
8	21	7.61	78.62
9	24	8.7	87.32
10	35	12.68	100
Total	276	100	

Part 2: Summary Statistics for School Districts

Variable		Mean	Std.	Observations
Pupil-to-teacher ratio	Overall	15.220	3.574	N = 1494
	Between Counties		2.829	n = 276
	Within Counties		2.151	n/N = 5.41304
Expenditure per Pupil($1000)	Overall	6.427	1.745	N = 1494
	Between Counties		1.343	n = 276
	Within Counties		1.189	n/N = 5.41304
Dropout Rate(%)	Overall	8.635	2.975	N = 1494
	Between Counties		3.038	n = 276
	Within Counties		0.914	n/N = 5.41304

Source: NCES CCD 1990 Universe Survey, Finance Survey, Dropout-Completion Survey for Local Education Agency

5.2 Monte Carlo Results

5.2.1 No Endogeneity in the Data Generating Process

Previous studies of school resource contributions do not have endogeneity control for the location choice. The simple aggregation models are used for estimation based on the argument that errors related to endogeneity of school choice are reduced if data are aggregated to the county or higher level. Therefore, it is interesting to study the performance of the Ad Hoc Simple Aggregation Model vs. the Correct Aggregation Model under this simplest assumption, i.e., there is no endogeneity in the data generation process where families all have the same attitude toward the importance of the school inputs. Since the school's contribution is the same across families in both the utility function and the production function, all μ_i's in equations (5-1) and (5-2) disappear. This being assumed, the difference of estimates in the three models is solely due to the different modeling specifications about aggregation of resources. With a sample size of 2,000, we expect the Full Location Information Model to recover all the true coefficients, since it correctly specifies the possibility of choosing a school district and achieve a test score conditional on choosing the school districts. We compare the variation of the estimation results from all three models to the true values, and thus, we get an idea of which model performs well over multiple running experiments.

Table 5-2 shows the results of production function parameters when estimating the three models without endogeneity assumptions in both the data generation process and the estimation process. The utility function parameter estimations are listed in *Appendix A*. To illustrate the consistency in estimation of the three different modeling specifications, we repeat the estimation of all coefficients 50 times with different generated sets of Gumbel Errors in the location choice utility function and Standard Normal Errors in the test score production function. Not surprisingly, the Full Location Information Model gives the best mean estimates out of the three for the 50 replications, and the Correct Aggregation Model follows the Full Location Information Model

closely. The Ad Hoc Simple Aggregation Model gives a mean estimate for all variables that could at best return the true coefficients within two standard deviation, except for the variable Dropout Rate. This variable, which we have shown in the summary statistics in Part 2 of *Table 5-1*, is the one that experience little variance within counties. It is expected to see the Ad Hoc Simple Aggregation Model perform well for this variable since it is able to capture between counties variances for Dropout Rate. For variables that have some inter county differences such as the Pupil-to-teacher ratio and the Expenditure per Pupil, we find significant downward biases for the Simple Aggregation Model from these 50 replications. This model has a relatively large mean square error. Mean estimates of this model from the 50 trials is smaller in absolute values than the true coefficients. This fact doesn't provide supporting evidence for Hanushek, Rivkin, and Taylor (1996) [11]'s argument that aggregate models tend to overestimate the contribution of school resources to the test outcome, but this experiment indeed shows that Ad Hoc model provide biased estimates of school effectiveness. Meanwhile, we found the performance of the Correct Aggregation Model to be quite impressive; both the Full Location Information Model and the Correct Aggregation Model have a mean square error within the range of zero to 0.1. The impressive performance of the Correct Aggregation Model shows that we could correctly estimate the marginal contribution of schools to location choice/education using this model with a relatively small sample size even if the exact school district information is unavailable.

We also list sub-sampling county estimation results for the production function in *Table 5-3*. The sub-sampling estimation is conducted using the family's chosen county and a random drawn subset of counties from all counties to form a new choice set that contains a subset of the original choice set that families are faced with when migrating. As illustrated in the beginning of this section, this technique was first developed by McFadden in the 1970s [19] and was used extensively for discrete choice studies with a large choice set. In *Table 5-3* we can see that the production function parameters are almost identical to estimations of different sampling assumptions over 50 replications. This is not a strange finding considering the assumption about

no endogeneity in school selection in the data generating process since the sub sampling method (choosing a subset of counties as the choice set) only affects utility function estimations when there is no connection between the two. The results show that the utility function which focuses on parents when making location choice is not correlated with the child educational production function estimations. In *Appendix B,* we list the estimation for utility function parameters in the no endogeneity case. Readers can see slight changes in utility function parameters under different sub sampling procedure. In the next two sections, we assess the endogeneity assumption of school selection. It is interesting to see whether McFadden [19]'s Sub Sampling Technique also gives consistently unbiased results in the endogeneity control case as well.

Table 5-2 Monte Carlo Comparison of Estimation Results
for Major Production Function Parameters
Three Models (Sample Size:2000, Number of replications: 50, No endogeneity in DGP)

	name	Pupil-to-teacher ratio	Expenditure per pupil	Dropout rate
Full Location Information Model	TRUE	-0.500	1.000	-0.500
	Mean Est.	-0.5028	1.3815	-0.4178
	std err	(0.041)	(0.174)	(0.046)
	MSE	0.0016	0.1751	0.0089
Correct Aggregation Model	estimate	-0.5022	1.388	-0.4183
	std err	(0.040)	(0.171)	(0.048)
	MSE	0.0016	0.1791	0.0089
Ad hoc Model	estimate	-0.1049	0.0206	-0.5051
	std err	(0.074)	(0.446)	(0.075)
	MSE	0.1615	1.1541	0.0056

Table 5-3 Monte Carlo Sub Sampling
Estimation Results for Major Production Function Parameters
(Sample Size:2000, Number of replications: 50, No endogeneity in DGP)

	Counties	10	15	25	50	100
				Pupil-to-teacher ratio		
Full Location Information Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.5028	-0.5028	-0.5028	-0.5028	-0.5028
	Std	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)
	MSE	0.0016	0.0016	0.0016	0.0016	0.0016
Correct Aggregation Model	Estimate	-0.502	-0.5019	-0.5021	-0.5022	-0.5022
	Std	(0.041)	(0.041)	(0.041)	(0.041)	(0.040)
	MSE	0.0016	0.0016	0.0016	0.0016	0.0016
Ad Hoc Simple Aggregation Model	Estimate	-0.1049	-0.1049	-0.1049	-0.1049	-0.1049
	Std	(0.074)	(0.074)	(0.074)	(0.074)	(0.074)
	MSE	0.1615	0.1615	0.1615	0.1615	0.1615
				Expenditure Per Pupil		
Full Location Information Model	TRUE	1.000	1.000	1.000	1.000	1.000
	Estimate	1.3815	1.3815	1.3815	1.3815	1.3815
	Std	(0.174)	(0.174)	(0.174)	(0.174)	(0.174)
	MSE	0.1751	0.1751	0.1751	0.1751	0.1751
Correct Aggregation Model	Estimate	1.3889	1.3885	1.3883	1.3881	1.388
	Std	(0.172)	(0.171)	(0.171)	(0.170)	(0.171)
	MSE	0.1802	0.1797	0.1793	0.179	0.1791
Ad Hoc Simple Aggregation Model	Estimate	0.0206	0.0206	0.0206	0.0206	0.0206
	Std	(0.446)	(0.446)	(0.446)	(0.446)	(0.446)
	MSE	1.1541	1.1541	1.1541	1.1541	1.1541
				Dropout Rate		
Full Location Information Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.4178	-0.4178	-0.4178	-0.4178	-0.4178
	Std	(0.046)	(0.046)	(0.046)	(0.046)	(0.046)
	MSE	0.0089	0.0089	0.0089	0.0089	0.0089
Correct Aggregation Model	Estimate	-0.4183	-0.4183	-0.4182	-0.4183	-0.4183
	Std	(0.048)	(0.048)	(0.048)	(0.048)	(0.048)
	MSE	0.0089	0.0089	0.0089	0.0089	0.0089
Ad Hoc Simple Aggregation Model	Estimate	-0.5051	-0.5051	-0.5051	-0.5051	-0.5051
	Std	(0.075)	(0.075)	(0.075)	(0.075)	(0.075)
	MSE	0.0056	0.0056	0.0056	0.0056	0.0056

5.2.2 Discrete Levels of Endogeneity in the Data Generating Process

This section explores the fact that residential choices by the parents could be endogenous in the dynamic process of achieving a higher test score for a child. As suggested in earlier chapters of this paper, families can have different opinions about the importance of school districts or have different expectations about their children's future. Such unobserved preference factors could affect the children's educational performance and their choice of schools. We use the discrete factor approximation method suggested by Heckman (1981), and extended by Mroz and Guilkey (1992) [21], and Mroz (1999) [23] to control the unobserved heterogeneity in parent's preferences about schools. A detailed description of this method is listed in *Appendix A.* The main argument of the discrete factor method is that conditional on the unobserved preference factor for the family, the multinomial distribution of the originally inter-correlated errors is jointly independent. The joint distribution of Gumbel Errors from the utility function and the Normal Random Errors in the educational production function can be integrated over discrete points of the heterogeneity factor. Mroz (1999) [23] has demonstrated that the discrete factor approximation method, with less than ten points of support estimated for the distribution of the unobservable, can approximate a wide variety of distribution of the unknown factor. This is superior in performance to the Mixed Logit (McFadden and Train, 2000 [31]) by a large grid of starting values and a fitted number of supporting points. The traditional Instrumental Variable approach for the endogeneity problem requires an instrumental variable that is largely correlated with test outcome but not correlated with location choice. We find such an instrument very hard to obtain for this type of study due to the complexity of choices and concerns people consider in migrating.

First, let's suppose the world is divided into five different types of people, with preference of schools ranked from low to high. This differentiated "attitude" of parents toward the importance of schools will affect parental educational contributions to their children thus

generates different educational outcomes from their children. In the data generating process of the Monte Carlo, we randomly assign observations in our sample into these five categories of preferences and use five supporting heterogeneity points in the estimation to see if the models are powerful enough to retrieve the "true" parameters used in the data generating process. We choose 5 points of support for the heterogeneity factor since the main purpose of this Monte Carlo study is to demonstrate and compare the performance test score outcome models with different specification of location choices under aggregation assumptions (the fitting of different modeling specifications for more than 5 supporting points may be explored in future studies). There are in total 19 parameters to estimate in this study: 3 marginal school contribution parameters to the utility function, 3 marginal contributions to the utility from heterogeneity interacting with school qualities, the location choice of family from equation (5-1), 3 marginal effects of school resources to the test score outcome from equation (5-2), and 5 heterogeneity points with possibilities of observing each of the points. Notice that in the estimation of the location choice utility function in equation (5-1), we have the school quality indicators interacting with the heterogeneity factor so the marginal contribution of schools is a weighted average of the estimated coefficients over all five heterogeneity factors (see *Appendix A* for more details). For test outcome contribution, the estimated coefficient should recover the true coefficients if the model is specified correctly. We focus the analysis on the model's ability to recover the "true" contribution of school quality indicators to the test score outcome under endogeneity control.

To show the distribution of the performance of our models in this discrete heterogeneity case, we replicate the estimation of all utility function and production function parameters 50 times using different randomly generated errors. For each set of generated outcomes (this includes the optimal chosen school district and the test score outcome of students based on the school quality they chose), we use 5 sets of randomly generated starting values to estimate the parameters for 150 iterations, then choose the set of estimated parameters that gives the best likelihood function value and let it go for 500 iterations. The final outcome would be the

estimation result for the one out of 50 generated outcomes. *Table 5-4* shows the mean estimations for the major production function parameters together with the standard errors calculated from these 50 replications. We also list simple OLS without endogeneity control results from these 50 trials. Note that the 5 points discrete factor approximations method performs much better than OLS no endogeneity controls for both the Full Location Information Model and the Correct Aggregation Model. This means that the Discrete Factor Approximation Method performs adequately in the case that the preference of parents indeed affects the location choice of the family because it reflects the right distribution of the errors. In *Table 5-4*, we see that the Ad Hoc Simple Aggregation Model gives biased average estimates for marginal contribution of schools to the student test performance. From the 50 replications, the Ad Hoc Simple Aggregation Model has the largest mean square error among the three models meaning it gives the farthest from the true parameters for the estimation of marginal contribution of school qualities to the location choice and test score outcomes. This result supports our expectation that using aggregated school resources in modeling location will largely affect the fitting of the model and thus leads to incorrect estimation. The Correct Aggregation Model, though not as convenient as simple aggregation models, should be the best model specification to be used in location choice when exact location choice is not known.

The Full Location Information Model and the Correct Aggregation Model perform well for different subsampling technique implementations as well. In *Table 5-5*, we show the mean, standard deviation and mean square error results for all three models in 50 replications under various sub-sampling drawing assumptions. For the Full Location Information Model and Correct Aggregation Model, the estimations give consistent and generally unbiased estimation results for the three major quality indicators that affect test score outcome under all sub-sampling sizes. The mean square error of these two models is very consistent over different sub sampling sizes, and the MSE is usually ten times lower than that of the Ad Hoc Simple Aggregation Model. Interestingly for the Ad Hoc Simple Aggregation Model, when the sample size increases, the

mean square errors from the estimations of contributions of school to test score outcome increase for two of the quality indicators, Pupil-to-teacher ratio and the Dropout Rate. This trend does not exist in the utility function parameters estimations (shown in the *Appendix A*).

Table 5-4 Monte Carlo Comparison of Estimation Results
for Major Production Function Parameters:
Three models, OLS vs. Endogeneity Control
(Sample Size:2000, number of replications: 50, DGP Discrete 5 points)

			Pupil Teacher Ratio	Expenditure per Pupil	Dropout Rate
		True	-0.5	1	-0.5
Simple OLS	Full Location Information Model	Est	-0.207	-0.120	2.538
		Std	(0.277)	(0.346)	(0.835)
		MSE	0.161	1.373	9.916
	Correct Aggregation Model	Est	-0.229	3.109	0.291
		Std	(0.785)	(2.024)	(0.535)
		MSE	0.677	8.460	0.906
	Ad hoc Model	Est	-0.163	-0.107	2.556
		Std	(0.233)	(0.359)	(0.628)
		MSE	0.167	1.353	9.725
Discrete Factor 5 Points Support	Full Location Information Model	Est	**-0.495**	**0.999**	**-0.498**
		Std	(0.060)	(0.131)	(0.063)
		MSE	0.004	0.017	0.004
	Correct Aggregation Model	Est	**-0.486**	**0.969**	**-0.468**
		Std	(0.063)	(0.175)	(0.066)
		MSE	0.004	0.031	0.005
	Ad hoc Model	Est	-0.031	-1.819	-0.869
		Std	(0.043)	(0.215)	(0.051)
		MSE	0.222	7.994	0.139

Table 5-5 Monte Carlo Sub Sampling Estimation Results
for Major Production Function Parameters
(Sample Size:2000, Number of replications: 50, DGP Discrete 5 points)

	Counties	10	15	25	50	100
				Pupil Teacher Ratio		
Full Location Info. Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.580	-0.591	-0.594	-0.619	-0.640
	Std	(0.012)	(0.019)	(0.017)	(0.011)	(0.035)
	MSE	0.0065	0.0086	0.0091	0.0143	0.0208
Correct Aggregation Model	Estimate	-0.606	-0.607	-0.610	-0.616	-0.613
	Std	(0.024)	(0.029)	(0.022)	(0.020)	(0.025)
	MSE	0.0119	0.0123	0.0125	0.0139	0.0134
Ad hoc Model	Estimate	-0.049	-0.036	-0.041	-0.032	-0.034
	Std	(0.274)	(0.041)	(0.011)	(0.012)	(0.011)
	MSE	0.2748	0.222	0.2318	0.232	0.267
				Expenditure Per Pupil		
Full Location Info. Model	TRUE	1.000	1.000	1.000	1.000	1.000
	Estimate	1.113	1.131	1.140	1.272	1.321
	Std	(0.083)	(0.098)	(0.096)	(0.013)	(0.063)
	MSE	0.0195	0.0265	0.0285	0.0742	0.1068
Correct Aggregation Model	Estimate	0.934	0.913	0.916	0.914	0.906
	Std	(0.134)	(0.090)	(0.091)	(0.094)	(0.070)
	MSE	0.0223	0.0156	0.0152	0.0161	0.0138
Ad hoc Model	Estimate	-1.620	-1.646	-1.793	-1.795	-1.795
	Std	(0.678)	(0.389)	(0.240)	(0.238)	(0.239)
	MSE	6.034	7.2728	7.0989	7.0974	7.0976
				Dropout Rate		
Full Location Info. Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.5657	-0.5685	-0.571	-0.6014	-0.629
	Std	(0.016)	(0.019)	(0.020)	(0.006)	(0.028)
	MSE	0.0046	0.005	0.0054	0.0103	0.0174
Correct Aggregation Model	Estimate	-0.5109	-0.4997	-0.5038	-0.5	-0.4966
	Std	(0.037)	(0.031)	(0.030)	(0.035)	(0.028)
	MSE	0.0015	0.001	0.0009	0.0012	0.0008
Ad hoc Model	Estimate	-1.0326	-1.0206	-0.8091	-0.809	-0.809
	Std	(0.106)	(0.034)	(0.018)	(0.018)	(0.018)
	MSE	0.2948	0.2722	0.0958	0.0958	0.0958

5.2.3 Continuous Endogeneity in the Data Generating Process

In the real data generating process we could not assume that preferences of the parents fall into 5 discrete categories. In this section, we build on the basic structure of the three estimation models in the last section by assuming that the endogeneity factor is distributed standard normally in the sample. Specifically, in the data generating process we give each family a randomly generated standard normal error to create the location choice outcome and test score outcome. This is equivalent to a random parameters model that the marginal contribution of the schools to location choice is assumed to be random among families. We use discrete factor approximation method with five points of support to estimate the three models. We also experiment with different sub sampling assumptions for all three models in this continuous endogeneity scenario.

Table 5-6 presents the comparison of estimation results for the production function from 50 replications when the choice set contains all counties in the data generating process. Estimates for the utility function parameters are listed in *Table A-3* in *Appendix A.* The OLS specification of the three models inherently ignores the diversity of parents' opinions about the significance of school. Even though that is the case, the Full Location Information Model and Correct Aggregation Model still yields better estimates for Pupil-to-teacher ratio and Expenditure per Pupil than the Ad Hoc Simple Aggregation Model even when this important endogeneity assumption is ignored in the estimation. The Ad Hoc Simple Aggregation Model only outperforms the other two models for the dropout rate variable in the educational production function.

When the endogeneity of school choices is controlled for in the modeling framework (lower half section of *Table 5-6*), the superiority of the two correct models becomes apparent. From here we could see a similar pattern of the performance of the three models compared to the

discrete heterogeneity case presented above. Both the Full Location Information Model and the Correct Aggregation Model gives an average estimate of all the parameters very near the true parameters, while the ad hoc model could not even get the signs of the expenditure per pupil variable correct. The mean square error from the 50 replications of estimates for each variable (including both endogenous and exogenous variables) is at least two thirds smaller than the Ad Hoc Simple Aggregation Model. This experiment, together with the findings in the last section for the discrete heterogeneity case, has launched significant warnings about simple aggregation in the data for the estimation: the aggregation bias has serious specifications bias and is consistent over multiple error generating experiments.

The sub-sampling experiment results for production function parameters for the continuous endogeneity case over 10, 15, 25, 50 and 100 counties are shown in *Table 5-7.* Results for utility function parameters are in *Table B-4* in *Appendix B.* Following the discrete error case in the last section, we could draw two conclusions from this table: 1) the Correct Aggregation Model and the Full Location Information Model performs well with the sub sampling technique. Mean estimates for all three major school quality indicators can recover the true coefficients within two standard error for all sub sampling assumptions. The mean square error for these two models are much lower than that of the Ad Hoc Simple Aggregation Model. 2) Mean Square Error for the Ad Hoc Simple Aggregation Model is unstable and does not show a decreasing trend when sample size (number of counties in the choice set) increases, so that we cannot draw the conclusion that for this model, county choice of the families follows the IIA assumption.

The Monte Carlo study under different distribution assumptions of the unobserved preference factor proved the prominent performance of the “correct” aggregation model over simple aggregation model. This is an important finding. Since aggregation cannot be avoided most of the time for macro level studies, most researchers find aggregation to be biased but can hardly find a substitute for it. The Correct Aggregation Model presents a way to overcome the simple aggregation in the data when the tradeoff between bias and variance is hard to overcome.

The Correct Aggregation model has proved itself to be relatively accurate (providing a mean estimate very near the Exact Location Information model) but has lower level of variation compared with the Ad hoc Simple Aggregation Model. In the second half of this paper, we use both the Correct Aggregation Model and the Ad Hoc Simple Aggregation Model to estimate the contribution of schools to the test score outcome using real 1990 data in the United States and see how different the estimation from these two functional forms can be applied to the real world example. Additional to the three major school quality indicators used in the Monte Carlo experiment, we add important factors affecting location choice to the utility function estimation such as the expected income of the family, and the distance to each location. We use a value added approach for the child educational production function estimation and control for related family background and student preparations. The data used for the empirical study is listed in the next section along with the findings.

Table 5-6 Monte Carlo Comparison of Estimation Results
for Major Production Function Parameters:
Three models, OLS vs. Endogeneity Control
(Sample Size:2000, number of replications: 50, DGP standard normal)

			Pupil Teacher Ratio	Expenditure per Pupil	Dropout Rate
		True	-0.5	1	-0.5
Simple OLS	Full Location Information Model	Est	-0.207	-0.120	2.538
		Std	(0.277)	(0.346)	(0.835)
		MSE	0.161	1.373	9.916
	Correct Aggregation Model	Est	-0.229	3.109	0.291
		Std	(0.785)	(2.024)	(0.535)
		MSE	0.677	8.460	0.906
	Ad hoc Model	Est	-0.163	-0.107	2.556
		Std	(0.233)	(0.359)	(0.628)
		MSE	0.167	1.353	9.725
Discrete Factor 5 Points Support	Full Location Information Model	Est	**-0.495**	**0.999**	**-0.498**
		Std	(0.060)	(0.131)	(0.063)
		MSE	0.004	0.017	0.004
	Correct Aggregation Model	Est	**-0.486**	**0.969**	**-0.468**
		Std	(0.063)	(0.175)	(0.066)
		MSE	0.004	0.031	0.005
	Ad hoc Model	Est	-0.031	-1.819	-0.869
		Std	(0.043)	(0.215)	(0.051)
		MSE	0.222	7.994	0.139

Table 5-7 Monte Carlo Sub Sampling Estimation Results for Major Production Function Parameters
(Sample Size:2000, Number of replications: 50, DGP Standard Normal)

	Counties	10	15	25	50	100
				Pupil-to-teacher ratio		
Full Location Information Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.5859	-0.5965	-0.5989	-0.6232	-0.6098
	Std	(0.018)	(0.023)	(0.021)	(0.014)	(0.069)
	MSE	0.0076	0.0097	0.0101	0.0153	0.0152
Correct Aggregation Model	Estimate	-0.6231	-0.6321	-0.6417	-0.6661	-0.5536
	Std	(0.014)	(0.012)	(0.009)	(0.013)	(0.098)
	MSE	0.0153	0.0175	0.0201	0.0277	0.0093
Ad Hoc Simple Aggregation Model	Estimate	-0.6464	1.1277	-0.1491	-0.151	-0.538
	Std	(0.457)	(2.575)	(0.681)	(0.677)	(0.013)
	MSE	0.1605	7.0692	0.4318	0.4277	0.0016
				Expenditure Per Pupil		
Full Location Information Model	TRUE	1.000	1.000	1.000	1.000	1.000
	Estimate	1.1361	1.1698	1.1776	1.2836	1.261
	Std	(0.099)	(0.128)	(0.127)	(0.026)	(0.134)
	MSE	0.025	0.0398	0.0422	0.0809	0.08
Correct Aggregation Model	Estimate	1.3805	1.3815	1.3976	1.4296	1.207
	Std	(0.073)	(0.060)	(0.049)	(0.037)	(0.258)
	MSE	0.1484	0.1479	0.1597	0.1854	0.0873
Ad Hoc Simple Aggregation Model	Estimate	1.0094	0.342	0.363	0.3649	0.7631
	Std	(1.139)	(0.696)	(0.812)	(0.811)	(0.256)
	MSE	0.8649	0.7557	0.8448	0.8423	0.0999
				Dropout Rate		
Full Location Information Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.5695	-0.5764	-0.5797	-0.607	-0.6048
	Std	(0.019)	(0.025)	(0.027)	(0.013)	(0.055)
	MSE	0.0051	0.0063	0.0068	0.0115	0.013
Correct Aggregation Model	Estimate	-0.6207	-0.6261	-0.6319	-0.6549	-0.5696
	Std	(0.010)	(0.012)	(0.012)	(0.015)	(0.089)
	MSE	0.0146	0.016	0.0175	0.0241	0.0101
Ad Hoc Simple Aggregation Model	Estimate	-0.9721	-0.681	-0.3131	-0.3135	-0.8248
	Std	(0.176)	(0.581)	(0.851)	(0.850)	(0.037)
	MSE	0.2435	0.2579	0.5176	0.5164	0.1064

CHAPTER SIX

ESTIMATION AND RESULTS

Our sample from the Geocode version of NLSY79 dataset and its Child-Mother supplement in year 1994 contains 3618 children age 5 to 15 with their family background and student outcome information. The county of residence of the households is available to the researcher, but not the exact school districts chosen by the families. For empirically estimating the Full Location Information Model, we need to be able to link the school district characteristics to a test score achieved in that school district; this requirement cannot be fulfilled here. So that empirically we can only apply the Correct Aggregation Model Specification and the Ad Hoc Simple Aggregation Model Specification to estimate the school's effect. This chapter identifies the estimation results from both models and compares their results with or without selection controls.

We take each observed mother's year 1979 county of residence as exogenous starting location and look at their county of residence for 1994. When the NLSY survey of youth began in 1979, the surveyed women were between age 14 and 21. 1994 is the fourth time period of the NLSY supplementary survey of children of those 1979 youth survey women. In this year, a number of children born to the NLSY-Youth mothers came to school age. We argue the mothers' year 1979 county of residence is exogenous in the subsequent year 1994 location choice, since for most women, age 14 is a long time before the mothers give birth to a child, and most women stay in their birth place at age of 14 but should migrate when they are married and have children. For 1994, the previous year test score value should be the 1992 math score of these NLSY-Supplement survey children. We have 3576 children in the year 1994 NLSY survey that report a math score, while there are 2584 out of these 3576 that have reported their math score in year 1992.

Another argument raised by studies in this field concerns the endogeneity of the mother's work choice (Mroz, Liu, and Van der Claauw 2007 [9]). Looking at the expected wage offers from part-time, full-time jobs locally together with father's annual income, mothers could make a work choice decision (together with the school district choice decision) that could directly affect the educational production function of the child since a full-time work choice rather than staying at home for the mothers should greatly decrease the parental input of children's education in this family. Work choice of mothers is assumed to be exogenous here since the main purpose of this study is to question on the bias caused by the aggregation of school resources, under or no under endogeneity control. Our endogenous variables for the structural model include all three school district characters: Pupil-teacher ratio, Expenditure per pupil, and Dropout rate. Exogenous variables affecting the location choice of the family are the expected full-time and part-time wage offer for mother from the location, the expected earnings offer for the father, and distance between the mother's year 1979 location and each prospect locations. Exogenous variables affecting the educational production function, based on the endogenously chosen school district, is the mother's AFQT score, previous period PIAT math score, child's age, mother's full-time or part-time working choice, sex of the child, and an indicator for non-white race. Since the exact school district choice is not known to the researcher (but the county of resident is known), only the Correct Aggregation Model and the Ad Hoc simple aggregation model can be used in this empirical study. We compare estimates from these two models in both with or without selection control.

We also distinguish the utility function for the school choice of the single mother families and the father-present families. For the father present families, the fathers should look at the expected annual earnings from that location and the mothers look at the expected wage offer from the location, together with the school characteristics of the local schools to make a location choice. For the single mother families, however, mothers should look at the expected wage offer from the location, together with the school characteristics to make a location choice. Earnings offers

should not be in the specifications of the utility function simply because there is no father. In the 1994 NLSY sample we have, there are 2448 out of the 3618 children having a father (either a biological father or a step father), while there are 1170 children that have only mothers present in year 1994. We put in a dummy variable in the utility function to distinguish single mother families from the two-parent families, and specify the utility function parameters separately for both. More specifically, together with school district characters: Pupil-teacher ratio, Expenditure per pupil, and Dropout rate in Ω_{i,j_k}, Other location specific characteristics idiosyncratic to the family affecting the location choice are the expected full-time and part-time wage offer for mother from the location, the expected earnings offer for the father, and distance between the mother's year 1979 location and each prospect locations. Following the utility function specification, equation (3-5) in chapter 3:

$$\begin{aligned} V_{i,j_k} &= \alpha^1{}_1 Dist_{i,k} + \alpha^2{}_1 Exp\_wage\_pt_{i,k} + \alpha^3{}_1 Exp\_wage\_ft_{i,k} \\ &+ \alpha^4{}_1 Exp\_earn_{i,k} + \beta(\mu_i)\Omega_{i,j_k} + \varepsilon_{i,j_k} \end{aligned} \tag{6-1}$$

Where $\beta_1(\mu_i) = \beta_1^0 + \beta_1^1 \mu_i$,

ε_{i,j_k} follows i.i.d. extreme value distribution.

We put in a dummy variable D that indicates whether the observation belongs to a single mother family or two-parent family. D=0 if there is no father, D=1 if there is a father. When D=0, $Exp\_earn_{i,k}$ for the family in the above equation has a value zero. The utility function is specified in the following way with the dummy variable D:

$$\begin{aligned} V_{i,j_k} &= \alpha^1{}_1 Dist_{i,k}(1-D_i) + \alpha^2{}_1 Exp\_wage\_pt_{i,k}(1-D_i) \\ &+ \alpha^3{}_1 Exp\_wage\_ft_{i,k}(1-D_i) + \alpha^4{}_1 Exp\_earn_{i,k}(1-D_i) + \beta_1(\mu_i)\Omega_{i,j_k}(1-D_i) \\ &+ \alpha^1{}_2 Dist_{i,k}(D_i) + \alpha^2{}_2 Exp\_wage\_pt_{i,k}(D_i) \\ &+ \alpha^3{}_2 Exp\_wage\_ft_{i,k}(D_i) + \alpha^4{}_2 Exp\_earn_{i,k}(D_i) + \beta_2(\mu_i)\Omega_{i,j_k}(D_i) + \varepsilon_{i,j_k} \end{aligned} \tag{6-2}$$

For the Ad Hoc Simple Aggregation Model, parents are assumed to look at the county average school characteristics to make a county choice, while the other assumptions are the same as above. The utility function for the family using Ad Hoc Simple Aggregation Model is thus specified as the following ($\overline{\Omega}_{i,k}$ represents county average school characteristics):

$$
\begin{aligned}
V_{i,k} &= \alpha^1{}_1 Dist_{i,k}(1-D_i) + \alpha^2{}_1 Exp\_wage\_pt_{i,k}(1-D_i) \\
&+\alpha^3{}_1 Exp\_wage\_ft_{i,k}(1-D_i) + \alpha^4{}_1 Exp\_earn_{i,k}(1-D_i) + \beta_1(\mu_i)\overline{\Omega}_{i,k}(1-D_i) \\
&+\alpha^1{}_2 Dist_{i,k}(D_i) + \alpha^2{}_2 Exp\_wage\_pt_{i,k}(D_i) \\
&+\alpha^3{}_2 Exp\_wage\_ft_{i,k}(D_i) + \alpha^4{}_2 Exp\_earn_{i,k}(D_i) + \beta_2(\mu_i)\overline{\Omega}_{i,k}(D_i) + \varepsilon_{i,j_k}
\end{aligned}
$$

(6-3)

Note that for both models, the marginal contribution of school characteristics to the possibility of families making a location choice is going to be differentiated for single mother families and the two-parent families. The marginal contributions from the expected wages and distance measures are also different. The first half portion of equation (6-2) and (6-3) represents the specification for the single parent families while second half represents the specifications of the two-parent families.

Another concern about this one-time-period estimation of the structural model using current period school and family inputs information is the control over the history of these two types of inputs. A popular remedy in analyzing cognitive educational achievement of children, when some of the past background information is missing, is to use a value added specification that assumes that a previous test score is a sufficient statistic for the missing historical inputs. Here, we follow Todd and Wolpin (2003) [32]'s work and implement a cumulative production function for children's cognitive achievement at a given age to depend on previous time period test score as well. For 1994, the previous year test score value should be the year 1992 PIAT math score of these NLSY-Supplement survey children. As mentioned in *Chapter 4* of this paper, we miss the year 1994 PIAT math score data for 2584 out of the 3160 children in the sample,

mostly because the children are not at school age in year 1986. But for these 1161 children we still have their location of residence information with all family background and school inputs, so we model the utility function for the families of the 2584 children making a school choice, but not the production function for educational outcome. Specifically, recall the Correct Aggregation Model specification of the likelihood function for each observation in equation (3-13) in *Chapter 3*:

$$L_i = \sum_{h=1}^{H} \Pr(\mu_{h,i}) \{ \sum_{j_{k^*}=1}^{J_{k^*}} \Pr(j_{k^*,i} = j_{k^*,i} \mid \Omega_{j_{k^*},i}, Z_{i,j_k}, \mu_{h,i}) \Pr(test_i = test_i^* \mid \Omega_{j_{k^*},i}, X_i, \mu_{h,i}) \}$$

Exogenous variables in X_i affecting the educational production function, based on the endogenously chosen school district, is the mother's AFQT score, previous period PIAT math score, a dummy showing whether the family moves from the mother's 1979 location, child's age, mother's full-time or part-time working choice, indicator variable for boy, and an indicator variable for non-white race. If this observation lacks previous period math score information, the likelihood function for this observation will be specified as the following with $\Pr(test_i = test_i^* \mid \Omega_{j_{k^*},i}, X_i, \mu_{h,i})$ equal to 1:

$$L_i = \sum_{h=1}^{H} \Pr(\mu_{h,i}) \{ \sum_{j_{k^*}=1}^{J_{k^*}} \Pr(j_{k^*,i} = j_{k^*,i} \mid \Omega_{j_{k^*},i}, Z_{i,j_k}, \mu_{h,i}) \} \tag{6-4}$$

This way we could still use the inputs from both the family side and the school side to estimate the utility function. Similar is true for the Ad Hoc Simple Aggregation Model. Recall the likelihood function specification for the Ad Hoc Simple aggregation Model using the county average school characteristics, equation (3-18):

$$L_i = \sum_{h=1}^{H} \Pr(\mu_h) \{ \Pr(k_i = k_i^* \mid \bar{\Omega}_{k,i}, Z_{i,k}, \mu_{h,i}) \Pr(test_i = test_i^* \mid \bar{\Omega}_{k^*,i}, Z_{i,k}, \mu_{h,i}) \}$$

If this observation lacks the previous period math score information, then the likelihood for this observation will be specified as:

$$L_i = \sum_{h=1}^{H} \Pr(\mu_h)\{\Pr(k_i = k_i^* \mid \bar{\Omega}_{k,i}, Z_{i,k}, \mu_{h,i})\} \tag{6-5}$$

For a complex structural likelihood function with more than 30 parameters like this one, a global maximum is hard to find. If starting from different starting values, different estimation of the models with or without endogeneity control may yield very different results, since it is possible that local maximum rather than global maximum is achieved in the optimization process. In this study, we start with the same set of starting values for each of the parameters to be estimated in the maximum likelihood function with estimation of all models to be compared (this includes Ad Hoc Simple Aggregation Model and Correct Aggregation Model with and without endogeneity control) The results presented here are the one set of estimation of all parameters with the highest likelihood function value from different trial starting values set. This section is organized as following: results of estimations without endogeneity are presented in section 6-1. Point estimates of both models with selection and endogeneity control and with different sub sampling assumptions, is listed in section 6-2.

6.1 Point Estimates of Production Function Parameters---No Endogeneity Control

Table 6-1 lists the point estimate of the production function parameters with the same set of starting values for the log likelihood function when no endogeneity control method (simple OLS and Standard Conditional Logit) is used. The reported log likelihood function value at the optimum is -33022.242 for the ad hoc Simple Aggregation Model and -33116.782 for the Correct Aggregation Model. A comparison of the Ad Hoc Simple Aggregation model and the Correct Aggregation Model shows that both models give a point estimate that conforms to the intuition for most of the variables in the school district characteristics and the family background (student preparation) factors. We find that without endogeneity control, the empirical estimation of the

two aggregation models yield very similar results. The marginal contribution of pupil teacher ratio to the PIAT math score goes from -0.015 in the ad hoc simple aggregation model to -0.016 in the correct aggregation model, which is only a 6% increase. The marginal contribution of expenditure per pupil and the drop out rate experiences a higher change from the simple aggregation to correct aggregation. The expenditure per pupil has a marginal contribution of 0.055 to the test score in the simple aggregation model. This means a 1000 dollar more input per student can improve the student test score by 0.055. The marginal contribution of this factor shows 0.076 in the correct aggregation model, which is also a trivial marginal contribution. Same notion is true for the drop out rate, both models show that a 100% percent drop out rate increase is going to decrease the student test score by 2.16 (simple aggregation model) to 3.15. This is not a small magnitude considering the maximum PIAT score is at a 10, but people can hardly imagine a school's dropout rate would double in the short run due to some policy change. This result, however, does not support Hanushek, Rivkin, and Taylor (1996) [11]'s argument that the aggregated data seems to augment school effect in child educational production.

Even though the estimates for the Correct Aggregation Model are generally slightly larger in magnitude than the Ad Hoc Simple Aggregation Model, we find that the estimates on the other family ground and student preparation factors are very similar to each other. Having a mother with higher AFQT score, who is working either part time or full time generally helps with the student's academic achievement; previous period test score also contributes positively to the current academic performance. Math scores also increases when children grow older with more academic sensitivity----being a boy helps, too. Last but not least, having a father in the household generally helps slightly with the children's academic performance, fathers contribute 0.05-0.06 point to their children's test score, which is a moderate level of improvement.

Table 6-1 Comparison of Estimation Results: OLS for Correct Aggregation and Ad Hoc Simple Aggregation Model
(No heterogeneity or selection control)

		Correct Aggregation Model		Ad Hoc Simple Aggregation Model	
	Variable	Estimate	Std. Err.	Estimate	Std. Err.
Production Function Parameters (Standard OLS)	Endogeneous Variables				
	Pupil Teacher Ratio	**-0.016**	(0.001)	**-0.015**	(0.014)
	Expenditure per Pupil	**0.076**	(0.016)	**0.055**	(0.040)
	Drop out Rate	**-3.150**	(0.018)	**-2.16**	(1.882)
	Exogenous Variables				
	Constant	1.566	(0.038)	1.575	(0.454)
	Mother's AFQT	2.045	(0.018)	2.053	(0.199)
	Previous Period PIAT Math Score	0.489	(0.021)	0.490	(0.018)
	Child Age	0.0322	(0.010)	0.033	(0.020)
	Mother Working Part-time	0.150	(0.195)	0.147	(0.119)
	Mother Working Full-time	0.209	(0.018)	0.205	(0.111)
	boy	0.150	(0.088)	0.148	(0.087)
	Non_white	-0.362	(0.091)	-0.368	(0.104)
	Father	0.061	(0.098)	0.056	(0.097)
	Std. of the error Term	2.190	(0.036)	2.193	(0.039)
	lnF	-33116.782		-33022.242	

6.2 Point Estimates of Production Function Parameters---With Endogeneity Control

Monte Carlo results from the last section show strong superiority of endogeneity control models over the simple OLS/logit models (Comparison shown in *Table 4-4* and *Table 4-6*). Endogeneity control models facilitate the assumption that the families are differentiated in opinions about importance of school. This section presents the results of estimating both Ad Hoc Simple Aggregation Model and the Correct Aggregation Model with selection control of school choices. With a five point discrete factor approximation of the preference factor, we estimate the marginal contribution of school quality indicators to both the possibility of parents choosing the school and the possibility that the child achieve a certain test score. We use five discrete points for controlling for the heterogeneity factor because Mroz (1999) [23] showed that five to eight heterogeneity support points will improve estimation results (in terms of likelihood function value). Since the heterogeneity factor is only differentiated among families but not from different school districts, we have it interacted with the school quality indicators in writing the utility function, so that the average contribution of schools to location choice is simulated across all five categories of this endogenous factor. Readers can find the technical details about the simulated marginal contribution of school districts to location choice procedures and results in *Appendix A.*

A unique characteristic of the endogeneity control model is that together with estimating the contributions of various school characters and student background factors, we also estimate the marginal contributions of the heterogeneity factor ('preference' of the family toward school choice) together with the location of the five point supporting points and their possibilities. This causes the precautions we make in the estimation. For a simple OLS/Logit, the distribution of the errors is assumed to be concave, so that a global maximum is easy to obtain using the numerical first derivative optimization program. In our case, we try to allow for an extremely flexible distribution of the error terms by relaxing the strict distributional assumptions. The nolinear optimization program sometimes stops at a local, rather than global maximum. To try our best to

avoid the local maximum, we randomly choose 5 sets of starting points, and use the numerical optimization program to run for 1000 iterations, until the first partial derivatives of each of our 41 parameters are adequately small---to the 6 to 7 decimal place after 0. We pick up the one set of parameter estimates that yield the best maximum likelihood value. With that being said, it could still possible we fail to grasp the global maximum: it's a tradeoff question that should be left for future discussion and research.

Table 6-2 shows the estimation results of the structural model with endogeneity control. at the maximum likelihood for the endogeneity case, the Ad Hoc Simple Aggregation Model has a likelihood function value of -32969.60 while the correct aggregation model has likelihood function value of -33070.837. Notice that both of the likelihood function values are higher than that of the no heterogeneity control case, meaning the heterogeneity control method fits the data better.

An obvious result from comparing *Table 6-2* with *Table 6-1*, i.e comparison of both models with or without endogeneity control, is that the magnitude for almost all the estimated endogenous variables becomes smaller with selection control. For the Ad Hoc Simple Aggregation Model, Parameters for Pupil-to-teacher ratio goes down from -0.015 to -0.004 when controlling for endogeneity. Estimated school quality contributions are similar to the no endogeneity control case for variable Expenditure per Pupil. For the correct aggregation model, The Dropout Rate parameter decreases sharply in terms of marginal contribution to test score from -3.15 to -1.82 for the Correct Aggregation Model, which is a 70% decrease in marginal contribution. the parameter estimate for Pupil-to-teacher ratio drops from -0.016 to -0.011. A reason this endogeneity control model behaves so differently from the no-heterogeneity control case is that it separated the effect of parents opinion about the importance of schools from the school input section, and parents are allowed to differentiate in terms of the preference they have for schools compared with other location specific characters thus the unobserved parental input to the child production function is thus controlled in the model.

Compared with the no heterogeneity control case, parameter estimates of other factors of the family background and location specific characteristics are differentiated between the two models in the endogeneity control case. The Correct Aggregation Model seems to give smaller in scale estimates for the family background factors, student preparation factors and gene factors. Interestingly, both models in the endogeneity control case gives an arbitrary sign of the heterogeneity factor (see parameter estimate for "Pho") to the test score, and a weakly negative contribution of "father presents" to the test score--- this probably means that the preference factors interact with the "father presents" dummy to affect the student outcome. The reason the full-time working status of mother has a positive contribution to educational production is because it is treated as an exogenuous variable here, and working mothers usually have better IQ rates that passed onto their children.

At the first glance, readers would find that the marginal contribution of the drop out rate is counter intuitive in the ad hoc simple aggregation model, for this heterogeneity control case. This conforms to our Monte Carlo studies that the simple aggregation of resources can yield biased or wrong-signed results. The ad hoc model doesn't give wrong signed estimate for the dropout rate in the OLS estimates but gives a wrong-signed estimate when parents' preferences are allowed to differentiated, this means that simple aggregation doesn't work well with the heterogeneity control scenario. A horizontal comparison of the Ad Hoc Simple Aggregation Model with the Correct Aggregation Model shows that the marginal contribution of two out of three endogenous school characters are smaller in magnitude if resources are aggregated "correctly". Marginal impact of Expenditure per Pupil drops from 0.081 to 0.078 when correctly modeling the incentive of school choice. The marginal impact from the local dropout rate greatly decreases when the Correct Aggregation Model is used. The marginal impact drops from 7.819 to -1.820. *Appendix C* details the utility function parameter findings.

Table 6-2 Comparison of Production Function Estimation Results: Discrete Factor Approximation for Correct Aggregation and Ad Hoc Simple Aggregation Model

		Correct Aggregation Model		Ad Hoc Simple Aggregation Model	
	Variable	Estimate	Std. Err.	Estimate	Std. Err.
Production Function Parameters (With Endogeneity Control)	Endogeneous Variables				
	Pupil Teacher Ratio	**-0.011**	**(0.012)**	**-0.004**	**(0.015)**
	Expenditure per Pupil	**0.078**	**(0.042)**	**0.081**	**(0.047)**
	Drop out Rate	**-1.820**	**(0.690)**	**7.819**	**(1.766)**
	Exogenous Variables				
	Constant	4.222	(0.367)	0.631	(0.376)
	Mother's AFQT	1.531	(0.168)	1.660	(0.172)
	Previous Period PIAT Math Score	0.361	(0.015)	0.394	(0.016)
	Child Age	0.031	(0.017)	0.042	(0.017)
	Mother Working Part-time	0.091	(0.098)	0.135	(0.103)
	Mother Working Full-time	0.155	(0.093)	0.152	(0.095)
	boy	0.117	(0.074)	0.118	(0.076)
	Non_white	-0.294	(0.089)	-0.285	(0.090)
	Father	-0.038	(0.083)	-0.028	(0.086)
	Pho	-3.571	(0.076)	-12.656	(5.565)
	Std. of the error term	1.357	(0.031)	1.233	(0.062)
	lnF	-32929.600		-33070.837	

CHAPTER SEVEN

CONCLUSION

Almost all of the models estimated in the literature of child educational production function are based on the assumption that schooling quality is exogeneously determined. To address this matter, Heckman et al. (1996) [13] clearly asserted "given the well-established link between family background and residential location decisions, this assumption is questionable." Therefore, we provide a flexible behavioral modeling framework here to facilitate the endogeneity control of the school selection problem by allowing parents to diversify in terms of their opinions (or unobserved parental inputs) toward the significance of educational production of their children. This fact has been taken into account when families make collective decision of location of residence and the education of their children. Because it reinforces the family's role in affecting the turnover of school resources into human capital, the model avails the conventional argument that children tend to become who their parents want and are. Compared with the parents' role in both the school choice and the everyday education of their children, the school's resources or other environmental impacts the children receive are of minor significance. Interpretation of the decreased impact from schools, when selection control is in place, is the "standard endogeneity of inputs argument that children with better family background and higher income sort themselves into better school district" (Heckman (1996) [13]), and at the same time, be more diligent in studying and working. The advantage of students in better school district disappears when the endogeneity of school resources has been accounted for in the structural model.

To help explain the empirical finding in the literature that estimates of quality effects derived from micro educational production functions based on aggregated schooling quality data tend to find stronger effects than studies using school-level or district-level quality data, we

compare two modeling specifications of location choice and educational production function under aggregation assumptions. The first model simply specifies the incentives of migration for families are based on the aggregated resource level and links the aggregated resources of their county of residence to the test score outcome of their children. The second model, which we call the "Correct Aggregation Model", accurately specifies the utility of parents choosing the location based on the best affordable school district of their county of residence, test score outcome is then linked probabilistically to that exact school resource their children receive. The Monte Carlo simulations of the two models yield the conclusion that the Correct Aggregation model could provide the most efficient and unbiased estimations of school effect while simply aggregating the resources to the county level will cause strong bias caused by the wrong specification of incentives. The empirical estimation of the two models yields the same conclusion as Hanushek, Rivkin, and Taylor (1996) [11] and Heckman (1996) [13]'s findings that aggregate models exaggerate school's effect on child educational production. Even when family background and sorting behavior of families across different locations are controlled, we have two major school quality indicators being strongly over-estimated by the Ad Hoc Simple Aggregation model. Our modeling specification (the Correct Aggregation Model) advances the literature by accurately specifying the incentives of parents sorting through school districts, even under data limitations about the exact school resource the child is receiving. Other than the specification error reasoning proved here, several authors argue that there may be measurement error in the inputs that is averaged out in aggregates (Card and Krueger (1994) [5], Betts (1995) [2], Hanushek, Rivkin, and Taylor (1996) [11]). Averaging the quality data could also substantially increase sampling variability of estimated quality effects by eliminating true variation in the data (Heckman 1996 [13]). Our Monte Carlo and empirical experiments support these findings.

APPENDICES

Appendix A: Simulated Marginal Contribution of School Districts to the Utility Function

We use discrete Factor Approximation Method (Mroz (1999) [23]) to estimate the structural model when endogeneity of school choice is in place. In this appendix we illustrate on how we parameterize the points of support and the probabilities for the discrete factor approximation estimation of the structural model. And describe the simulation of marginal contribution of school district to the utility function of the family. We majorly follow Mroz (1999) [23] for the convenient normalizations, details about the whole process and experiments can be found in his paper.

Suppose we have h points of support for μ of each family. It is possible for μ to fall on any of the h point μ_h. We restrict the μ_h's to lie on [0,1], with μ_1 =0 and μ_H =1. For this purpose, for the values of points in between we use standard logit to obtain the sub optimization over the h-2 parameters $\theta_2, ..., \theta_{H-1}$:

$$\mu_h = \frac{\exp(\theta_h)}{1+\exp(\theta_h)}, h = \theta_2, ..., \theta_{H-1}$$

For the probability of each point of support for μ_i, we consider the constraint that those h probabilities have to sum up to 1. Also each probability has to be non negative. We use an easy parameterization as the following to satisfy these two constraints:

Define

$$\tau_h = 1 + \sin(\frac{3\pi}{2} + \theta_h) \quad h = 1, 2, ..., H-1$$

$$\tau_h = 1 + \sin(\sum_{h'=1}^{H-1} \theta_{h'}) \quad h = H$$

And let $P_h = \frac{\tau_h}{\sum_{h'=1}^{H-1} \tau_{h'}}$.

In the maximum likelihood estimation with 5 points discrete support, we obtain 3 estimations for μ_h=2,3,4 and 4 estimations of P_h=1,2,3,4. According to equation (3-5), for the utility function we have:

$$V_{i,j_k*} = \alpha_1 Z_{i,j_k} + \alpha_2(\mu_i)\Omega_{i,j_k*} + \varepsilon_{i,j_k*} \quad (3\text{-}5)$$

Where $\alpha_2(\mu_i) = \alpha_2^0 + \alpha_2^1 \mu_i$

The marginal contribution of Ω_{i,j_k*} to the utility function should be an average of $\alpha_2(\mu_i)$ across all points of support for the unobserved discrete factor. For example, in estimating the Full Location Information Model with 5 points of heterogeneity support in the Monte Carlo, I got the set of estimation as following in the first run:

Table A-1 First Run 5 Point Discrete Factor Approximation for the Full Location Information Model in Monte Carlo

NAME	TRUE	Estimates	STD
X 1	-0.50	-.496***	(0.068)
X 2	1.00	1.003***	(0.051)
X 3	-0.50	-.495***	(0.604)
X_MIU 1	-0.50	-.465***	(0.092)
X_MIU 2	1.00	.9719***	(0.014)
X_MIU 3	-0.50	-.450***	(0.007)
Loading on 1st to 5th order discrete factors with probabilities			
HETERO(1	0.00	0	0
P(1)	0.20	0.0256	0
HETERO(2	0.40	0.3796	0
P(2)	0.20	0.0256	0
HETERO(3	0.60	0.7295	0
P(3)	0.20	0.2582	0
HETERO(4	0.80	0.732	0
P(4)	0.20	0.3162	0
HETERO(5	1.00	1	0
P(5)	0.20	0.3742	0

Then the marginal contribution of X1 to the utility function is calculated by:

$$X1+\sum_{i=1}^{5} X\_MIU_i * HETERO(i) * P(i)$$

Similarly is true for X2 and X3.

Table A-2 lists the utility function parameters estimations for the three models for 50 replications. We have similar results for production function estimations: the Correct Aggregation Model follows the Full Location Information Model most closely and yield an estimation that easily recovers the true coefficients in the data generation process. We also list comparison of estimation results for utility function parameters in *Table A-3* and *Table A-4*, when the unobserved preference factor μ_i is distributed discretely and standard normally. From the tables we see that the performance of Correct Aggregation Model over the Ad Hoc Simple Aggregation Model to be very promising. The comparison of simple Logit and discrete factor approximations results also show the superiority of endogeneity control models over no endogeneity control models.

Table A-2 Monte Carlo Comparison of Estimation Results for Utility Function Parameters: Three Models (Sample Size:2000, Number of replications: 50, No endogeneity in DGP)

	name	Pupil-to-teacher ratio	Expenditure per pupil	Dropout rate
Full Location Information Model	TRUE	-0.500	1.000	-0.500
	Mean Est.	-0.559	0.9234	-0.5236
	std err	(0.031)	(0.145)	(0.026)
	MSE	0.0044	0.0266	0.0012
Correct Aggregation Model	estimate	-0.5597	0.9217	-0.5238
	std err	(0.032)	(0.149)	(0.027)
	MSE	0.0046	0.0279	0.0013
Ad hoc Model	estimate	-0.3321	0.3985	-0.4427
	std err	(0.015)	(0.056)	(0.019)
	MSE	0.0284	0.3649	0.0036

Table A-3 Monte Carlo Comparison of Estimation Results for Utility Function Parameters: Three Models (Sample Size:2000, Number of replications: 50, Discrete 5 Points in DGP)

			Pupil-to-teacher ratio	Expenditure per Pupil	Dropout Rate
	True		-0.78	1.56	-0.78
Simple OLS	Full Location Information Model	Est	-0.046	0.902	-0.631
		Std	(0.204)	(0.051)	(0.083)
		MSE	0.247	0.012	0.024
	Correct Aggregation Model	Est	-0.444	0.328	-0.203
		Std	(0.886)	(1.782)	(0.335)
		MSE	0.773	3.564	0.198
	Ad Hoc Simple Aggregation Model	Est	0.004	0.896	-0.674
		Std	(0.218)	(0.036)	(0.069)
		MSE	0.301	0.012	0.035
Discrete Factor 5 Points Support	Full Location Information Model	Est	**-0.877**	**1.829**	**-0.910**
		Std	(0.028)	(0.050)	(0.022)
		MSE	0.010	0.075	0.017
	Correct Aggregation Model	Est	**-0.878**	**1.830**	**-0.911**
		Std	(0.028)	(0.049)	(0.023)
		MSE	0.010	0.075	0.018
	Ad Hoc Simple Aggregation Model	Est	-2.232	2.703	-0.491
		Std	(0.113)	(0.413)	(0.084)
		MSE	2.121	1.473	0.090

Table A-4 Monte Carlo Comparison of Estimation Results for Utility Function Parameters: Three Models (Sample Size:2000, Number of replications: 50, Standard Normal Endogeneity in DGP)

			Pupil-to-teacher ratio	Expenditure per Pupil	Dropout Rate
		True	-0.78	1.56	-0.78
Simple OLS	Full Location Information Model	Est	-0.086	0.915	-0.612
		Std	(0.221)	(0.049)	(0.070)
		MSE	0.220	0.010	0.017
	Correct Aggregation Model	Est	-0.010	0.899	-0.667
		Std	(0.233)	(0.037)	(0.068)
		MSE	0.295	0.012	0.032
	Ad Hoc Simple Aggregation Model	Est	-0.523	0.222	-0.210
		Std	(0.946)	(1.781)	(0.309)
		MSE	0.894	3.775	0.180
Discrete Factor 5 Points Support	Full Location Information Model	Est	**-1.165**	**2.219**	**-0.997**
		Std	(1.305)	(2.521)	(0.980)
		MSE	1.817	6.663	0.989
	Correct Aggregation Model	Est	**-0.664**	**1.205**	**-0.657**
		Std	(1.296)	(2.641)	(1.053)
		MSE	1.659	6.960	1.103
	Ad Hoc Simple Aggregation Model	Est	-1.937	2.670	-0.353
		Std	(0.325)	(0.398)	(0.091)
		MSE	1.443	1.387	0.190

Appendix B: Sub Sampling Estimation Results

See tables *B-1* to *B-4.*

Table B-1 Monte Carlo Sub Sampling Estimation for Major Utility Function Parameters, no endogeneity in the DGP, Estimation without Endogeneity Control

	Counties	10	15	25	50	100
				Pupil-to-teacher ratio		
Full Location Information Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.5459	-0.5557	-0.5458	-0.5473	-0.5476
	Std	(0.034)	(0.061)	(0.031)	(0.031)	(0.031)
	MSE	0.0032	0.0067	0.003	0.0032	0.0032
Correct Aggregation Model	Estimate	-0.5465	-0.5567	-0.5465	-0.5482	-0.5487
	Std	(0.037)	(0.062)	(0.033)	(0.033)	(0.032)
	MSE	0.0035	0.0069	0.0032	0.0034	0.0034
Ad Hoc Simple Aggregation Model	Estimate	-0.3886	-0.4992	-0.366	-0.3489	-0.3369
	Std	(0.024)	(0.047)	(0.020)	(0.018)	(0.016)
	MSE	0.013	0.0021	0.0149	0.0232	0.0268
				Expenditure Per Pupil		
Full Location Information Model	TRUE	1.000	1.000	1.000	1.000	1.000
	Estimate	0.9026	0.8995	0.899	0.898	0.8978
	Std	(0.155)	(0.156)	(0.156)	(0.156)	(0.154)
	MSE	0.0331	0.0339	0.034	0.0342	0.0337
Correct Aggregation Model	Estimate	0.9015	0.8971	0.8962	0.8939	0.8932
	Std	(0.159)	(0.160)	(0.159)	(0.158)	(0.156)
	MSE	0.0344	0.0356	0.0355	0.0359	0.0353
Ad Hoc Simple Aggregation Model	Estimate	0.5001	0.4873	0.47	0.44	0.4122
	Std	(0.106)	(0.100)	(0.086)	(0.070)	(0.064)
	MSE	0.2609	0.2726	0.2726	0.3183	0.3495
				Dropout Rate		
Full Location Information Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.4178	-0.5555	-0.5547	-0.555	-0.5549
	Std	(0.046)	(0.062)	(0.058)	(0.056)	(0.057)
	MSE	0.0089	0.0068	0.0063	0.0061	0.0062
Correct Aggregation Model	Estimate	-0.4183	-0.5565	-0.5553	-0.5554	-0.5552
	Std	(0.048)	(0.062)	(0.058)	(0.056)	(0.057)
	MSE	0.0089	0.007	0.0064	0.0062	0.0062
Ad Hoc Simple Aggregation Model	Estimate	-0.5051	-0.4945	-0.4869	-0.4765	-0.4672
	Std	(0.075)	(0.045)	(0.040)	(0.038)	(0.036)
	MSE	0.0056	0.002	0.002	0.002	0.0023

Table B-2 Monte Carlo Sub Sampling Estimation for Major Utility Function Parameters, Discrete Five Points Endogeneity in DGP, Discrete Five Points Approximation Estimation

	Counties	10	15	25	50	100
				Pupil-to-teacher ratio		
Full Location Information Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.2927	-0.3033	-0.3054	-0.3233	-0.336
	Std	(0.128)	(0.141)	(0.165)	(0.171)	(0.186)
	MSE	0.2531	0.2465	0.2514	0.2368	0.2306
Correct Aggregation Model	Estimate	-0.3117	-0.3269	-0.3273	-0.339	-0.3512
	Std	(0.156)	(0.178)	(0.150)	(0.166)	(0.166)
	MSE	0.2428	0.236	0.2266	0.221	0.2106
Ad Hoc Simple Aggregation Model	Estimate	-0.1411	-0.2234	-0.2726	-0.2088	-0.2141
	Std	(0.435)	(0.114)	(0.181)	(0.081)	(0.088)
	MSE	0.5907	0.3223	0.289	0.3326	0.3276
				Expenditure per pupil		
Full Location Information Model	TRUE	1.000	1.000	1.000	1.000	1.000
	Estimate	0.0941	0.1164	0.1027	0.0956	0.1109
	Std	(0.085)	(0.117)	(0.100)	(0.083)	(0.083)
	MSE	2.1557	2.0973	2.1334	2.1509	2.1066
Correct Aggregation Model	Estimate	0.1082	0.0606	0.0839	0.0809	0.0988
	Std	(0.090)	(0.101)	(0.071)	(0.064)	(0.083)
	MSE	2.1155	2.2579	2.1838	2.1918	2.1417
Ad Hoc Simple Aggregation Model	Estimate	-0.0519	-0.0087	0.0095	-0.0035	-0.0105
	Std	(0.286)	(0.197)	(0.185)	(0.126)	(0.136)
	MSE	2.6771	2.498	2.4371	2.4601	2.4842
				Dropout rate		
Full Location Information Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.3708	-0.3898	-0.3898	-0.413	-0.428
	Std	(0.091)	(0.098)	(0.116)	(0.135)	(0.146)
	MSE	0.1754	0.1616	0.1652	0.1522	0.1443
Correct Aggregation Model	Estimate	-0.4199	-0.4276	-0.4617	-0.4834	-0.4862
	Std	(0.092)	(0.105)	(0.107)	(0.118)	(0.110)
	MSE	0.1378	0.1348	0.1123	0.1015	0.0979
Ad Hoc Simple Aggregation Model	Estimate	-0.4685	-0.3654	-0.3776	-0.345	-0.3346
	Std	(0.337)	(0.110)	(0.128)	(0.103)	(0.103)
	MSE	0.2069	0.1837	0.1777	0.1993	0.2086

Table B-3 Monte Carlo Sub Sampling Estimation for Major Utility Function Parameters, Standard Normal Continuous Endogeneity in DGP, Discrete Five Points Approximation Estimation.

	Counties	10	15	25	50	100
				Pupil-to-teacher ratio		
Full Location Information Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.0982	-0.213	-0.3409	-0.6113	-0.3838
	Std	(0.215)	(0.240)	(0.435)	(0.073)	(0.054)
	MSE	0.1924	0.1209	0.1514	0.016	0.0155
Correct Aggregation Model	Estimate	-0.7237	-0.7154	-0.671	-0.5558	-0.582
	Std	(0.154)	(0.103)	(0.045)	(0.093)	(0.091)
	MSE	0.0659	0.0535	0.0306	0.0088	0.0122
Ad Hoc Simple Aggregation Model	Estimate	1.6264	1.1775	0.6544	0.4907	0.4427
	Std	(1.088)	(0.769)	(0.002)	(0.000)	(0.003)
	MSE	3.5994	1.5097	1.3326	0.9814	0.8888
				Expenditure per pupil		
Full Location Information Model	TRUE	1.000	1.000	1.000	1.000	1.000
	Estimate	0.2032	0.4594	0.7087	1.2602	0.7998
	Std	(0.489)	(0.526)	(0.925)	(0.158)	(0.136)
	MSE	0.7944	0.4767	0.6556	0.0844	0.0523
Correct Aggregation Model	Estimate	1.4477	1.3921	1.3109	1.0709	1.1535
	Std	(0.260)	(0.260)	(0.065)	(0.226)	(0.244)
	MSE	0.2454	0.1987	0.0994	0.0391	0.0631
Ad Hoc Simple Aggregation Model	Estimate	-10.2801	-2.0393	-1.5756	-1.1165	-1.6762
	Std	-9.9437	-2.6971	-1.9097	-1.3868	-0.0226
	MSE	17.3665	9.4259	9.0649	5.7617	7.1626
				Dropout rate		
Full Location Information Model	TRUE	-0.500	-0.500	-0.500	-0.500	-0.500
	Estimate	-0.1592	-0.2385	-0.3282	-0.5432	-0.3723
	Std	(0.142)	(0.152)	(0.298)	(0.068)	(0.079)
	MSE	0.1296	0.0837	0.0885	0.0049	0.0204
Correct Aggregation Model	Estimate	-0.6582	-0.6277	-0.5852	-0.4813	-0.5512
	Std	(0.092)	(0.112)	(0.078)	(0.047)	(0.061)
	MSE	0.0307	0.0246	0.0113	0.0018	0.0051
Ad Hoc Simple Aggregation Model	Estimate	2.3657	1.7705	3.0982	3.6467	5.1564
	Std	-1.6461	-2.0672	-4.9924	-4.8610	-1.4697
	MSE	7.6418	2.9005	2.9563	3.2948	3.3435

Table B-4 Sub Sampling Estimation Results for Major Production Function Parameters Using Year 1994 data:
Correct Aggregation and Ad Hoc Simple Aggregation Model
With 5 Points Heterogeneity Control

Sub Sampling Counties	Pupil-to-teacher ratio	Expenditure per Pupil	Dropout Rate
	Correct Aggregation Model		
25	0.009	0.061	-4.196
	(0.0160)	(0.0380)	(0.1800)
50	-0.016	0.052	-2.124
	(0.0120)	(0.0560)	(0.5700)
100	-0.016	0.052	-2.124
	(0.0152)	(0.0560)	(0.5700)
200	-0.017	0.077	-3.155
	(0.0125)	(0.0260)	(0.6600)
500	-0.017	0.077	-3.155
	(0.0126)	(0.0270)	(0.6770)
1000	-0.017	0.077	-3.155
	(0.0124)	(0.0524)	(0.6925)
	Ad Hoc Simple Aggregation Model		
25	-0.003	0.080	5.827
	(0.015)	(0.037)	(1.609)
50	-0.031	0.055	-1.486
	(0.013)	(0.036)	(0.272)
100	-0.031	0.055	-1.486
	(0.020)	(0.041)	(1.278)
200	-0.028	0.054	-1.769
	(0.017)	(0.040)	(1.113)
500	-0.015	0.055	-2.163
	(0.013)	(0.038)	(0.318)
1000	-0.015	0.055	-2.163
	(0.005)	(0.009)	(0.010)

Appendix C: Empirical Estimation for Major Utility Function Parameters—1994 Data

Our findings about the marginal contributions of the school characteristics to the location choice of the families are all mixed for both models using the 1994 data. Table C-1 lists the estimation results of the utility function parameters when no heterogeneity control is in place. We see the single parent families look at the location specific characteristics similarly from the two parent families. For the heterogeneity control case in the father presenting household, both the expected earnings for father and the expected full time wage offer for mother has a positive contribution to the possibility of the household choosing that location. Single mother families seem to put more emphasis on the expected wage for the mother to choose a location. The marginal effect of distance is also larger for single mother families than for the two parent families. This conforms to our intuition, since for single mother families it is more costly to move, also the expected earnings are very important since the mother is the only bread earner.

The heterogeneity control models yield roughly the same (but slightly different in magnitude) estimate about the location specific characteristics idiosyncratic to the family. With endogeneity control, the expected full time wage offer and the expected annual earnings for the father is still significantly positively affect the location choice decision. But the expected part time wage offer for the mother switches sign from no heterogeneity control to heterogeneity control models. This means the preference of the location interact with the women's decision of whether work full time or part time.

Table C-1 Estimation
for Major Utility Function Parameters for 1994 data
No Heterogeneity Control

		Correct Aggregation Model		Ad Hoc Simple Aggregation Model	
	Variable	Estimate	Std. Err.	Estimate	Std. Err.
Utility Function Parameters (Two Parents Family)	Endogeneous Variables				
	Pupil Teacher Ratio	0.1229	(0.005)	0.1948	(0.008)
	Expenditure per Pupil	-0.083	(0.017)	0.097	(0.021)
	Drop out Rate	10.94	(0.088)	2.256	(0.967)
	Exogenous Variables				
	Expected Fulltime Wage for Mother	1.937	(0.038)	3.635	(0.308)
	Expected Earnings for Father	1.144	(0.028)	1.358	(0.129)
	Distance	-0.013	(0.479)	-0.046	(0.039)
	Expected Parttime Wage for Mother	-0.201	(0.018)	-0.055	(0.297)
Utility Function Parameters (Single Mother Family)	Endogeneous Variables				
	Pupil Teacher Ratio	0.1273	(0.008)	0.2006	(0.011)
	Expenditure per Pupil	0.0009	(0.014)	0.162	(0.028)
	Drop out Rate	15.29	(0.012)	14.36	(1.128)
	Exogenous Variables				
	Expected Fulltime Wage for Mother	2.095	(0.428)	4.044	(0.238)
	Distance	-0.523	(0.182)	-0.632	(0.192)
	Expected Partime Wage for Mother	-0.1	(0.057)	-0.098	(0.058)

Table C-2 Estimation
for Major Utility Function Parameters for 1994 data
With 5 Points Heterogeneity Control

		Correct Aggregation Model		Ad Hoc Simple Aggregation Model	
	Variable	Estimate	Std. Err.	Estimate	Std. Err.
Utility Function Parameters (Single Mother Family)	Endogeneous Variables				
	Pupil Teacher Ratio	0.123	(0.015)	0.164	(0.019)
	Expenditure per Pupil	0.015	(0.036)	0.098	(0.055)
	Drop out Rate	14.955	(0.571)	-37.905	(2.185)
	HET*Pupil Teacher Ratio	0.001	(0.022)	0.046	(2.007)
	HET*Expenditure per Pupil	-0.057	(0.060)	0.071	(0.881)
	HET*Drop out Rate	0.937	(0.656)	66.082	(14.607)
	Exogenous Variables				
	Expected Fulltime Wage for Mothe	0.680	(0.315)	2.998	(0.434)
	Distance	-0.118	(0.770)	-0.116	(0.683)
	Expected Partime Wage for Mother	1.633	(0.558)	1.151	(0.057)
	Simulated Marginal contribution				
	Pupil Teacher Ratio	0.124		0.155	
	Expenditure per Pupil	-0.012		-0.035	
	Drop out Rate	15.390		-31.517	
Utility Function Parameters (Two Parents Family)	Endogeneous Variables				
	Pupil Teacher Ratio	0.118	(0.010)	0.152	(0.013)
	Expenditure per Pupil	-0.095	(0.035)	-0.045	(0.042)
	Drop out Rate	9.797	(0.576)	-34.100	(1.764)
	HET*Pupil Teacher Ratio	0.010	(0.016)	0.058	(16.050)
	HET*Expenditure per Pupil	0.021	(0.055)	0.188	(32.527)
	HET*Drop out Rate	2.222	(0.672)	47.696	(34.628)
	Exogenous Variables				
	Expected Fulltime Wage for Mothe	1.633	(0.288)	1.168	(0.438)
	Expected Partime Wage for Mother	1.940	(0.267)	3.635	(0.307)
	Expected Earnings for Father	1.144	(0.122)	1.357	(0.130)
	Distance	-0.013	(0.038)	-0.048	(0.039)
	Simulated Marginal contribution				
	Pupil Teacher Ratio	0.123		0.166	
	Expenditure per Pupil	-0.085		0.101	
	Drop out Rate	10.830		-34.327	

Appendix D: Defining the 1987-2005 US School Districts

D.1 Overview

Common Core data files (CCD) from the National Center for Education Statistics (NCES) provide more than 20 years of school level, school district level and state level local education agency information about almost all aspects for local schooling environment across the United States. It has three major survey categories for the local school districts level data: Local Education Agency (School District) Universe Survey Data: 1986–Present; Local Education Agency (School District) Finance Survey (F-33) Data: 1990–Present; Local Education Agency (School District) Universe Survey Dropout and Completion Data: 1991–Present. For every local school district agency, there is a consistent ID code called LEAID (Local Education Agency ID) assigned by NCES to the agency, the first two position of the LEAID code identifies the FIPS state code for the local agency. This dataset is very helpful for researchers concerned with local educational achievement, budget, expenditures, local housing market and local labor market supply issues in the United States.

We found some difficulties when using this data resource, however, to obtain a panel data of the local school agency quality information nationwide from 1986 to present. Firstly, the years of survey conducted in the three categories of survey experience some inconsistency. The universe survey covers from school year 1986 to 2006 with little changes in its contents and variables of interests, while Finance Survey covers only year 1990, 1992, 1995 to present; the Dropout Survey dates back to year 1991, but has up to 80% of agency characteristics missing for some years. Variables that have strong missing problems in the Dropout Survey are the per grade dropout base, per grade dropout count. This makes integrating the three dataset and obtaining a comprehensive longitudinal dataset hard to fulfill. Secondly, there are new school districts come into existence or an old local agency is merged, closed, or transferred into another type of agency during the 25 year period. CCD provides variables such as BOUND (dummy for change of

boundary) and STATUS (Indicator for "active" or not) to identify this information but again, these two variables are not available for each year. Moreover, there is no provided information of previous name or LEAID of agency if it disappears from the datasets. Thirdly, the CCD system is designed to be "inclusive" rather than "exclusive". Thus, CCD files contain a substantial number of records representing administrative and operating units that are unlike typical public schools and school districts—for example, there are schools or districts without students and special education schools for disabled or American Indians. There are 7 types of schools for some years of CCD Universe Survey, these types range from regular school to special education school, vocational school and schools for minorities, disabled etc. This makes it hard to distinguish local administrative agencies that report exclusively, or partially, from the "real" local school agencies that has teachers and students, class buildings or geographic standings. Last but not least, it is hard to find a representative schooling environment for a multiple children family if the local agency is only a primary or secondary school district that does not offer a degree up to 12. The limited geographic information provided by the CCD (ZIPCODE, CITY NAME, and FIPS) causes us a lot of trouble integrating local agencies to be one that provide degrees of education from kindergarten up to high school diploma that is available to the local residents.

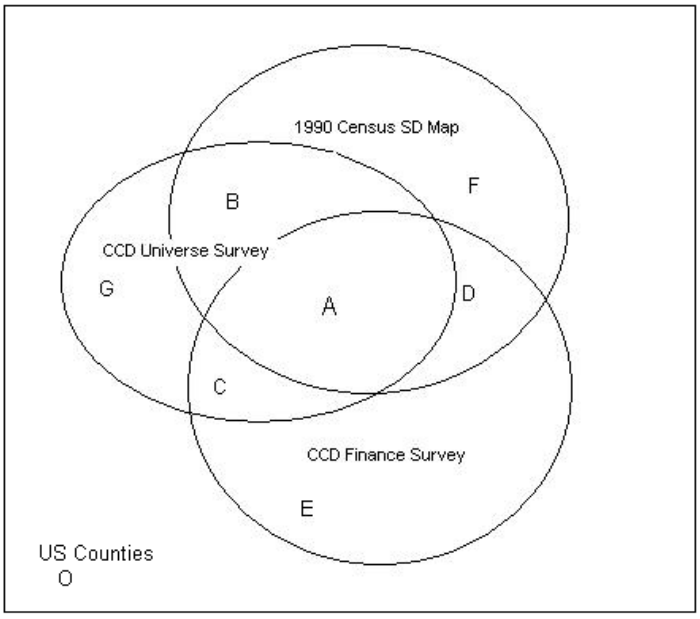
Our purpose here is to obtain a longitudinal dataset that has information for the school districts that have been in existence in the United States, for all the years from 1987 to 2005. Our Variables of Interest are: number of students, number of teachers, teacher salary, total revenue, total expenditure, and total dropout rate. For this purpose of well defining the local education agencies that are distinguished from the administrative agencies or vocational services that are for the adults, and having degree offers up to 12, we need to know very clearly the geographic occupation of the school districts. We use the 1990 Census School District Map (referred to as "Map" in the later part of this appendix) to serve this purpose. The Map has 15512 local education agencies, 12920 of them are defined as the "Unified" or primary/secondary school districts that offers degree up to high school diploma. 3582 of them are "Standalone" school

districts where only partial degrees (not up to 12) are offered. For each of the 3582 in the map, it is easy to find a nearby unified school district a certain student might go to within the county by Latitude and Longitude geographic information.

D.2 Data Files

Graph C-1 shows the data files we use and the overlapping among them. As mentioned in the last section, we use the 1990 Census School District Map as our key reference for the definition of 1990 school districts. The other three files we use are CCD Universe Survey for Local School District Agency, 1987-2005; CCD Finance Survey for Local School District Agency, 1990, 1992, 1995-2005; CCD Dropout Survey for Local School District agency, 1991-2005; and Ipums3171 US County file. The last file is downloaded from the Geographic Tools web page of IPUMS.org, it is a comprehensive datafile that have all 3170 1990 US counties with boundary files, and PUMA (Public Use Microdata Area), METAREA (Metropolitan Area Status) information, county, and state FIPS information.

Graph D-1 Overlapping of Map, CCD Universe, and CCD Finance Survey



We start by defining an effective 1990 school districts file for the US. We take year 1990 Universe Survey and Finance Survey to be merged with the 1990 Census School District Map. This gives us 15017 school districts (LEAIDs) that is in area A of Graph C-1. Note that the Dropout survey is left out in the first steps here for two reasons: firstly, the dropout survey was not conducted in year 1990, it started at year 1991; secondly, for every year of dropout survey we found serious missing data problem. Of roughly 15000 school districts that joined the survey each year, only about 2000 report dropout base and dropout count information. So we put aside this survey until next section when we come back and fill in the values for each of our defined school district using geographic neighboring or year neighboring interpolations/extrapolations strategy, to obtain the dropout data from this survey.

As mentioned in the last section, the Census School District Map is our key reference in defining "real" school districts that distinguished from administrative agencies or others, so for each LEAID in the map, we know it is a reasonable school district but the agencies happened to be left out from the survey in year 1990, or it has a kind of joint reporting prototype with some other nearby agency. The possibility of joint reporting is also one of the reasons that we do not use ZIPCODE as the main reference for definition, since it turns out that several agencies geographically far from each other may use one service agency to implement their general/financial businesses, so that in CCD the several school districts share one ZIPCODE which is misleading. Therefore, for the LEAIDs in the map but not in either of CCD universe or finance (Area F in Graph C-1), we use their closest neighboring LEAID's survey data as the "real" information for that LEAID but retain the map LEAID of the agency.

The areas G+C in graph D-1 represents LEAIDs that exists in the 1990 CCD Universe (and may or may not in the CCD Finance Survey) but not identified by the 1990 Census School District Map. We drop TYPE 3,4,5,6 (TYPE 3,4,5,6 stands for administrative agencies or vocational agencies that are not typical school districts) from this group of school districts, since they are not common education agency according to the document from NCES. For the TYPE 1

and TYPE 2 leftover districts (roughly 329 of them) that took part in the Universe Survey but not in the map, we use ZIPCODE and CITY information to merge them into the map, treating them as the same agent as the LEAIDs with same ZIPCODE in the map (zip2scldst.dta, resulted document ccd90-matched.dta). We conduct this task by adding in the data of each to-be-merged agent to the merged agent. One thing worth noting here is that for area E, the 88 LEAIDs took part in finance survey but not in either the Universe Survey or in the Map in year 1990, we ignore them, since they do not provide any other information than LEAID, so it is not possible for us to affiliate them with some "motherhood" agency or nearby agency.

There are certain school districts that span two counties in our database. We have the availability of inter-district school choices for residents in the two adjacent counties that are in the attendance boundary of these school districts. We identify this type of school district by the difference in the COUNTY ID provided by the Map and the CNTYCODE provided by the CCD Universe Survey. For those LEAIDs providing different county information from the map and from CCD Universe Survey, we treat them as existing in both counties (that is to treat them as two different LEAIDs, available to residents in both of the counties). We still have 23 counties that is not in the Map nor CCD (universe and finance) but in our 3170 standard county dataset (ipums3171.dta), we use "overlap" county/school district file which is a byproduct from the Census School District Map to identify 11 more of them (leaid2cty.dta), each identified LEAID is put into one of the 11 counties as a separate school district existing both of the counties. We therefore have in hand a school district file for the year 1990 that has 14765 school districts for the 3152 counties in the United States: the 28 other counties are mostly counties like "Yellowstone National Park" that do not virtually have school districts in them, we could thus safely ignore those counties.

Another problem is the degree types that are offered in these school districts. If we could not make sure each of the identified school district offers degree up to 12^{th}, it would add to the complication of our research to identify the school district of each child of a multiple-children

family. Also a majority of the school districts in the US has degree range from kindergarten to high school diploma. We need to guarantee that all the school information for a typical school-children is being collected in the datasets. Therefore, for all the LEAIDs that do not offer a degree up to 12 (standalone school districts), we use their neighboring LEAIDs that offer degree up to 12 information on top of their own information (mainid.dta, subidsv2.dta, leaidall.dta). For example, suppose there is a district A that is a standalone where nearby there is a school district B that offers full degree. District B (full degree)'s data might go to two different places -- one providing information to be grouped with district A(standalone), and the other for its own separate existence as a school district. If B has elementary school students as well as older grades, we would be counting the elementary students (and older students) in B as if they were in district A (they would also be in B as that is where they really are).

Up to this step we have a data file for the school districts in the United States, with their area of occupation, district boundaries, degrees offered, and other related information well defined and identified for the year 1990. We could thus hold this file as our standard school district file and obtain each of their information or change of their characteristics in other years from 1987 till 2005. We redo the whole process that listed in section I and section II for each year from 1987 to 2005 other than 1990, holding 14765 LEAIDs in the year 1990 as our Static School District standard file. If an LEAID exist in our 1990 file do not appear in another year, say 1992, then we leave the position for that LEAID, we will come back to this LEAID in section III to fill in its 1992 information using multi-year interpolation/extrapolation. Details about this strategy will be discussed in Section III.

D.3 Interpolation and Extrapolation

D.3.1 Dropout Survey

We did a little more work than normally expected, to the all year's dropout survey data. Problems associated with the CCD dropout survey is that, firstly, it started in year 1991, a year

later than we define our 1990 standard school districts. So it is necessary to fill in year 1990 information using other available years. Secondly, all years of the dropout survey have a serious missing information problem. This will greatly decrease our sample size if we use the raw data to merge into the full sample. So we use a two step data filling in strategy here that starting with a typical year with certain dropout information the original dataset provides, keep our filling using the LEAID's own other grade reporting, or state average per grade drop rate information, for every year. In the second step we use linear interpolation across years to fill in values that are missed out in the first step. The filling follows the following rules that are common in the literature about data linear interpolation:

Step 1:

Define the dropout rate of a school district (R_k) as the ratio of sum of degree 7 to degree 12 drop count (D) and degree 7 to degree 12 Enrollment Base (E):

$$R_k = \sum_{i=7}^{12} D_i \Big/ \sum_{j=7}^{12} E_j$$

Time subscripts are missed here for notation convenience. Rewrite it as:

$$R_j = \frac{1}{\sum_{i=7}^{12} E_i} \sum_{j=7}^{12} E_j \left[\frac{D_j}{E_j} \right] \qquad \text{(C-1)}$$

Now define the average enrollment base as the sum of available enrollment base for each grade divided by the number of grades that have enrollment base available:

$$N_k = \sum_{available} E_i$$

$$avgN_k = \sum_{available} E_i \Big/ n_{available}$$

We calculate equation (C-1) under certain data filling rule. If E_j is available, then we have E_j as in the formula, if it is not available, then we use $avgN_k$ to represent the value for the missing E_j.

Similarly, if $\frac{D_j}{E_j}$ is not available, we use the state average per grade dropout rate to fill in that value. Note the state level per year per grade dropout rate is calculated as:

$$st\_avgR_k = \sum_{available\_k} D_{available\_k} \Big/ E_{available\_k}$$

This recovers a larger portion of the dropout survey for the 1990 static school district, but we still do not have dropout ratio for each LEAID in each year. Remember in Step 1 we have only used cross section information (state dropout rate) and per grade dropout rate to recover the data, in the next step we want to explore the longitudinal information (cross years) to recover almost all the dropout survey data for all years.

Step 2:

In filling the missing year (the year that has reported no dropout base) dropout ratio for a typical LEAID in between two other available years, we consider both the two available years' dropout ratio, together with the number of grades that have "dropbase" data for the two available years. More specifically, in calculating the weighted average, we put more weights into the year that have more dropout base information, since this year is indeed more "reliable". Suppose we have in the dataset for a LEAID that miss year 1995 dropout ratio, but have 1994 and 1996 dropout rate available, while year 1994 it has n_{94} (ranges from 1 to 6) grades reported dropout base and in 1996 it has n_{96} (ranges from 1 to 6) grades reported dropout base. Then we use the following equation to interpolate the dropout rate for this LEAID in year 1995 (notice that R_{95} might not be zero, according to the calculations in Step 1):

$$R_{95} = \frac{(1+n_{94})*R_{94} + (1+n_{96})*R_{96} + R_{95}}{(1+n_{94}) + (1+n_{96}) + 1}$$

Suppose we have two years in between that has missing dropout information. We use state average dropout rate on top of longitudinal filling, because we do not want to totally rely on the information provided by a single agency. For example, if we have an agency reporting no dropout base data for year 1995 and 1996, but do have some information for year 1994 and 1997, and suppose for 1994 the number of grades that provide dropout base is n_{94} and 1997 this number is n_{97}. Then we go through 2 steps to get a good enough simulated result for dropout rate in 1995 and 1996.

$$tR_{95} = \frac{\frac{2}{3}(1+n_{94})*R_{94} + \frac{1}{3}(1+n_{97})*R_{97}}{\frac{2}{3}(1+n_{94}) + \frac{1}{3}(1+n_{97})}$$, and

$$R_{95} = \frac{[(1+n_{94})+(1+n_{97})]*tR_{95} + st\_avgR_{95}}{(1+n_{94})+(1+n_{97})+1}$$

Similarly for year 1996:

$$tR_{96} = \frac{\frac{1}{3}(1+n_{94})*R_{94} + \frac{2}{3}(1+n_{97})*R_{97}}{\frac{1}{3}(1+n_{94}) + \frac{2}{3}(1+n_{97})}$$

$$R_{96} = \frac{[(1+n_{94})+(1+n_{97})]*tR_{96} + st\_avgR_{96}}{(1+n_{94})+(1+n_{97})+1}$$

More generally, for any missing year YR, let's suppose the first non-missing year before year YR is BG_YR with droprate R_{BG_YR}, n_{BG_YR} valid dropout base and the first non-missing year after year YR is ED_YR with drop rate R_{ED_YR} and n_{ED_YR} valid dropout base for grades from 7 to 12, then we firstly calculate the temporary simulated-dropout rate as following:

$$tR_{YR} = \frac{\frac{BG\_YR - YR}{ED\_YR - BG\_YR}(1 + n_{BG\_YR}) * R_{BG\_YR} + \frac{BG\_YR - YR}{ED\_YR - BG\_YR}(1 + n_{ED\_YR}) * R_{ED\_YR}}{\frac{BG\_YR - YR}{ED\_YR - BG\_YR}(1 + n_{BG\_YR}) + \frac{BG\_YR - YR}{ED\_YR - BG\_YR}(1 + n_{ED\_YR})}$$

Then we obtain the simulated dropout rate by using the state average dropout rate and the calculated temporary dropout rate from above:

$$R_{YR} = \frac{[(1 + n_{BG\_YR}) + (1 + n_{ED\_YR})] * tR_{YR} + st\_avgR_{YR}}{(1 + n_{BG\_YR}) + (1 + n_{ED\_YR}) + 1}$$

Now, we have dealt with missing years in between two valid years, but remember we have in the dataset years from 1987 till 2005. The first dropout survey year is 1991, so we miss at least 4 years before the first year, and for a majority of LEAIDs we possibly miss the tails year too. To fill in the dropout rate values for the "headers and footers", we define the year with missing information as YR, the first year after YR that has valid information to be BG_YR, then if we define T as T=YR-BG_YR-1, then the years we need to fill in value for can be indexed as BG_YR-t, t goes from 1 to T. The following extrapolation strategy make sure the weights put into R_{BG_YR} is higher the nearer it is to the missing year, and weights put into the state average dropout rate to be higher when the BG_YR is far away from the missing year YR :

$$R_{YR} = R_{BG\_YR-t} = (\frac{1}{t+1})R_{BG\_YR} + (1 - \frac{1}{t+1})R_{st\_avg\_YR}$$

After this, if we still have years for a LEAID that miss information, then it is probably the case that there is no dropout base information for this agency at all in the dropout survey, we have to give them the state average dropout rate.

D.3.2 Universe/Finance Survey

The data filling for other years (than 1990) in Universe/Finance survey of local education agency is comparably easier, since agencies typically report student number, teacher number, and annual financial information in these two surveys. We do not have serious missing data problems

in all the years when survey is available; typically we have more than 85% percent positive values for our interested variables for the 14765 “static” school districts each year. The reason for us to do some data filling for these two surveys is that, firstly, for finance survey we have all agency information missing for year 1987, 1990, 1989, 1991, 1993 and 1994 when the survey is not conducted, holding the fact that school district financing would not change dramatically for the same LEAID over years, we can easily obtain the missing year information from other years by interpolation, unless this LEAID is closed. Secondly, for Universe Survey we do have less than 10% of the agencies reporting zero and/or missing student number, to obtain our ratio variables such as the pupil/teacher ratio, expenditure per pupil, or teacher salary per pupil, we could not leave the student number or teacher number to be zero for an LEAID at certain year while a year before and after that this number is not zero.

Holding the fact that local education agency would not experience dramatic changes in terms of student/teacher number, annual revenue/expenditure in total and teacher salary in total, we conduct simple interpolation/extrapolation to the missing data in both surveys. We take weighted average of available year survey data for the same agency to represent missing year data. Suppose for year YR, the year before YR that has valid information to be BG_YR, the year after YR that has valid information is ED_YR, then if we define T as T=YR-BG_YR-1, then the years we need to fill in value for (YR) can be indexed as BG_YR-t, t goes from 1 to T.

$$Var_{YR} = Var_{BG\_YR+t} = \frac{t}{t+1}(Var_{ED\_YR}) + (1 - \frac{t}{t+1})(Var_{BG\_YR})$$

For headers and footers, since the teacher number and student number, expenditure, revenue and teacher salary for a school district do not vary too much over the years, so we keep them at the same real level with the last or first available year.

D.4 Wrap up

Recall that we have merged the 1990 Universe Survey to the 1990 Census School district map and merged the 1990 Finance Survey to the 1990 Census School District map, if a map LEAID appears in any of the two surveys, we put them into our 14765 static school district file. This will return the fact that if an LEAID never joined the finance survey, it is possible that it is in our 14765 static file with Universe Survey information but never a finance survey information, so all the data filling-in processes in Section III will not do any good but still give us the "missing" value for all finance information for the typical agency. The same will happen to the pupil/teacher variables to a LEAID if it never takes part in the Universe Survey but some years of Finance survey. Until this point we have recovered more than 99 percent of the dataset, but we still have roughly 20 locations where we have positive number of teachers/students but no finance at all for all years, or positive revenue/salary/expenditure but no student or teachers. This is truly not realistic, and to deal with it, we use the "nearest neighbor strategy" once again to find out a nearby agent's related information to represent the problem school district's information. Suppose an LEAID is lacking of all years finance data but having teacher/student number on file, after we find out a nearby agency that has full information to substitute for it, we drop the original teacher/student number of the problem agency, since it is no longer matched with the "true" finance data we give it.

To make sure we do not put zero values in the denominators, we check the variable "number of students" once more for reported zero number of students. The checking gives us only two agencies with zero student number. So we use interpolation /extrapolation onto these two agencies, obtain their positive report from other years, and assign the weighted average to them according to our data filling rule listed in Section III.

At last, we have 24 consistent years Universe/Finance/Dropout survey information for 14765"static" School District, that cross 3130 counties in the United States.

Appendix E: Expected Wage/Earnings Offer for the Parents

E.1 Defining the US Labor Market

Beginning July, 2007 we started the project of defining the US local labor markets based on geographic information from the Census/CPS geographic information. The purpose of this study is to get somewhat punctual identification of local labor employment region and the distance between them, in order to be used in the weighting scheme for the estimation of the average wages of specific age, sex, education, and race groups of people in a typical labor market. This article illustrates how our local labor markets are defined, and how we calculate the weighted geographic distance between those labor markets.

E.1.1 Geographic Data Source and Variables

IPUMS.org is the main source of geographic data we adopted. IPUMS.org developed geographical variables from the original Census/CPS database that can consistently identify locations in the database from 1977 through 2005. The variables that we use to define the local labor markets are:

- METAREA: Metropolitan Area, defined by Census/CPS as "counties or combinations of counties centering on a substantial urban area. METAREA identifies the metropolitan area where the household was enumerated, if that metropolitan area was large enough to meet confidentiality requirements." This variable is available over 1977 through 2005 and is consistent over Census/CPS database.
- CONSPUMA: Consistent Public Use Microdata Area, defined by IPUMS as "the most detailed geographic areas that can consistently be identified across samples from 1980 to 2006. It is an effort to reconcile differences in the lowest level geographical areas over years while retaining the highest possible level of geographic detail. This variable is consistent for 1980, 1990, 2000 Census data only.

- STATEFIP: State (FIPS code), defined by Census/CPS as the state in which the household was located, using the Federal Information Processing Standards (FIPS) coding scheme, which orders the states alphabetically. This variable is consistent over all years.
- CNTYGP98: County Group 1980, defined by IPUMS as an un-recoded variable that identifies the household's 1980 county group of enumeration. It is state-dependent; it must be read in conjunction with state codes, STATEICP and STATEFIP. This variable is available in year 1980 Census only, but experiences only slight changes into 1990, 2000 PUMA variables.

We based our geographic definition on the 1980 Census. As mentioned above, CNTYGP98 shows household's 1980 county group of enumeration, they are contiguous areas with a combined population of 100,000+ residents; they may consist of actual county groups, but they may also be single counties, single cities, or other census-designated places (separate or combined) meeting the 100,000+ population criterion. So we link counties into this variable using "Detailed Component of 1980 County Groups" file available in IPUMS (http://usa.ipums.org/usa/volii/ctygrp.shtml). Note that we also obtained information about counties that form a METAREA from this file.

CONSPUMA was defined as (STATEFIP*1000) + CNTYGP98 in this year, so we matched the above CNTYGP98 component file with the "CONSPUMA component file" available also in IPUMS (http://usa.ipums.org/usa/volii/conspuma_components.xls). This far we got a detailed dataset of all 3171 counties in the US, with information of the state, metropolitan area and CONSPUMA they belong to.

More effort can be taken to link counties into METAREAs and CONSPUMAs for year 1990 and 2000 Census using component files for 1990 and 2000 PUMAs. (http://usa.ipums.org/usa/volii/puma.shtml, and http://usa.ipums.org/usa/volii/2000pumas.shtml). We avoid this extra work in our labor market definitions since the matched files from 1980 can

identify more than 90% of the data. For those METAREAs enumerated after year 1980 that affects 10% of the whole surveyed population, we assigned them manually into nearby identifiable 1980 METAREAs depending on distance. The detailed reassignment is provided in *Table E-1*.

E.1.2 Local Labor Markets and Urban/Rural Centers

We use METAREA as the relevant labor market for counties that are "urban." (Those that are in a METAREA). We use 99+CONSPUMA+STATEFIP for the local labor market for counties that are non-urban – Note that we only know the CONSPUMA in 1980, 1990, and 2000 by using the Census data. We obtain in total 496 labor markets in the US, 272 of them are urban, and 274 of them are rural.

For the 274 rural areas where we use CONSPUMA as the labor market, we cannot link directly to information in the CPS in off census years. But we can link to state-wide rural locations. We use the overall state rural as part of the weighted average. Therefore we have 50 more "place holders" with a rural center in our dataset, but they are not defined as local labor markets that we will estimate wages for.

In order to get the most accurate distance information between labor markets, we define the center latitude/longitude for our local labor markets. The geographic latitude/longitude information for all counties in the US is obtained from the Census (http://www.census.gov/tiger/tms/gazetteer/county2k.txt). We take the 1980 population weighted average latitude/longitude of the counties that form a typical METAREA or CONSPUMA, and define that as the geographic center of the local labor market. To illustrate this, take part of South Carolina as an example (*Figure 1*), here Greenville County (point D), Spartanburg County (point E) and Anderson County (point F) form a metropolitan area, METAREA 3160: "Greenville-Spartanburg-Anderson SC", so we define an urban labor market 3160 and calculate the urban center by:

$$\mathrm{LAT}_{mean} = \sum_{i=1}^{3} \frac{pop_i}{tpop_{Metarea}} \times Latitude_i \text{ And } LONG_{mean} = \sum_{i=1}^{3} \frac{pop_i}{tpop_{Metarea}} \times Longitude_i$$

Figure D-1 shows a rural CONSPUMA, CONSPUMA 425 formed by Pickens County (point A) and Oconee County (point B) in state South Carolina (state 45). So we define a rural labor market by adding the rural indicator "99" onto the CONSPUMA and STATEFIP code, labor market 99425045 and getting the rural center (point C) using the functions listed above.

Similarly, the rural "Place holder" defined for CPS off years data in state South Carolina can be obtained by taking the geographic average of all rural counties in this state. (not shown in figure)

After getting the geographic center of all the local labor markets, the spherical distance between any two labor markets can be calculated from the function:

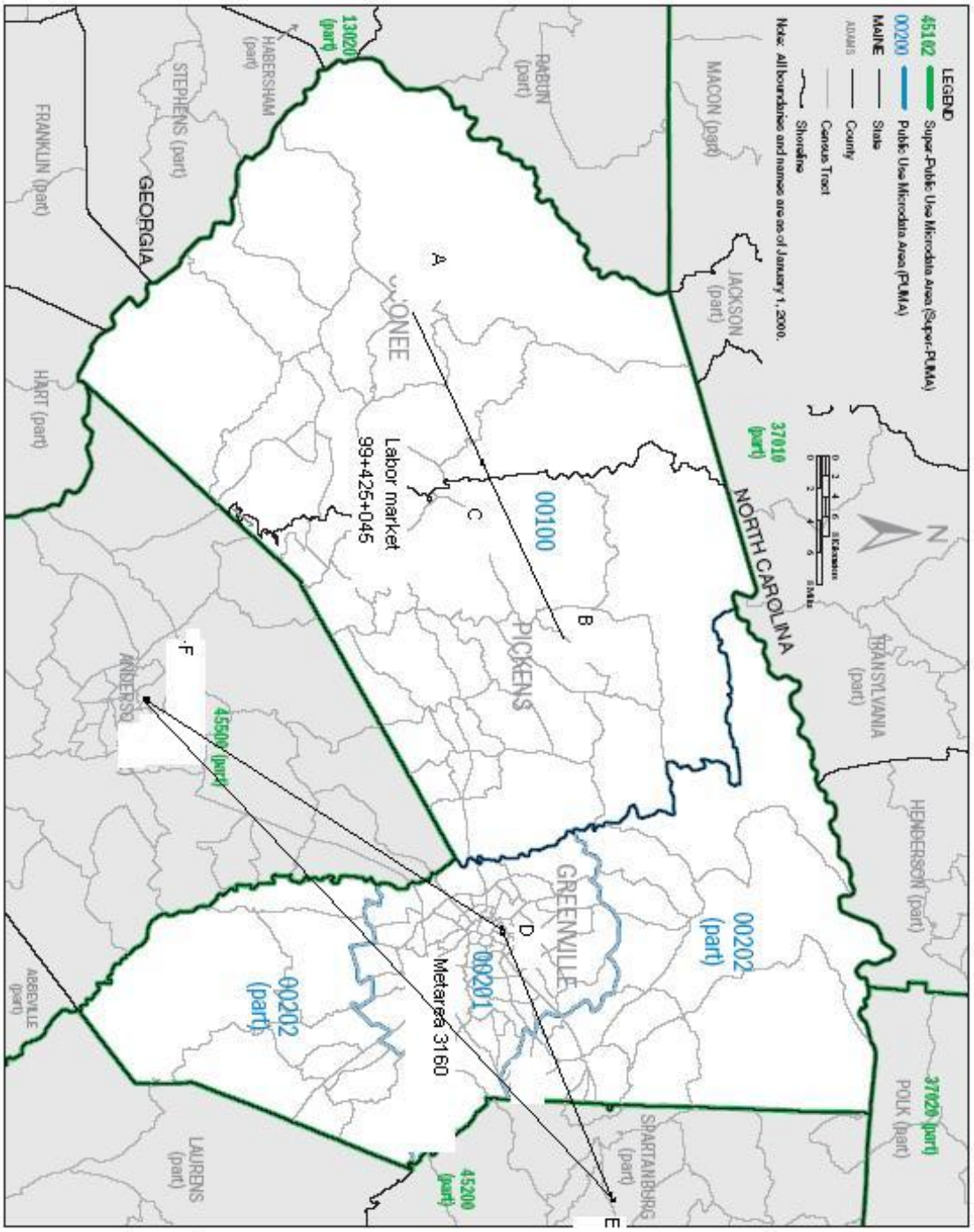
$$\Delta\sigma = \arctan\left(\frac{\sqrt{(\cos\phi_f \sin\Delta\lambda)^2 + (\cos\phi_s \sin\phi_f - \sin\phi_s \cos\phi_f \cos\Delta\lambda)^2}}{\sin\phi_s \sin\phi_f + \cos\phi_s \cos\phi_f \cos\Delta\lambda} \right)$$

$\phi_f, \lambda_f, \phi_s, \lambda_s$ being the geographical center latitude and longitude of two labor markets.

Table E-1 Merged METAREAs from Other Years into 1980 METAREAs

METAREA		Merged to METAREA	
480	Aseville,NC	1520	Charlotte-Gastonia-Rock Hill,NC
580	Auburn-Opekika, AL	450	Anniston, AL
740	Barnstable-Yarmouth, MA	1120	Boston, MA-NH
1020	Bloomington, IN	400	Anderson, IN
1123	Same name match-merging	1122	Same name match-merging
1200	Brockton, MA	1120	Boston, MA-NH
1260	Bryan-College Station, TX	1240	Brownsville-Harlingen-San Benito, TX
1281	Same name match-merging	1280	Same name match-merging
1300	Burlington, NC	1520	Charlotte-Gastonia-Rock Hill, NC-SC
1310	Burlington,VT	4760	Manchester, NH
1601	Aurora-Elgin	1600	Chicago-Gary-Lake,IL
1603	Joliet	1602	Chicago-Naperville-Joliet,IL
1604	Lake county into Chicago-Naperville-Joliet,IL	1600	Chicago-Naperville-Joliet,IL
1660	Clarksville- Hopkinsville, TN/KY	1740	Columbia, MO
1921	Fort Worth-Arlington	1920	Dallas-Fort Worth,TX
1930	Danbury, CT	1160	Bridgeport, CT
2030	Decatur, AL	1000	Birmingham, AL
2081	Boulder-Longmont	2080	Denver-Boulder-Longmont,CO
2180	Dothan, AL	1000	Birmingham, AL
2190	Dover, DE	6160	Philandelphia, PA
2280	Dutchess Co., NY	3810	Kileen-Temple, TX
2281	Dutchess County	3810	Kingston,NY
2520	Fargo-Morehead, ND/MN	2240	Duluth-Superior, MN/WI
2600	Fitchburg-Leominster, MA	1120	Boston, MA-NH
2620	Flagstaff, AZ-UT	6520	Provo-Orem, UT
2650	Florence, AL	1000	Birmingham, AL
2710	Fort Pierce, FL	2700	Fort Myers-Cape Coral, FL
2720	Fort Smith, AR/OK	2580	Fayetteville-Springdale, AR
2980	Goldsboro, NC	2970	Glens Falls, NY
3010	Grand Junction, CO	3060	Greeley, CO
3150	Greenville, NC	3120	Greensboro-Winston Salem-High Point, NC
3285	Hartford	8880	waterbury,CT
3300	Hattiesburg, MS	3560	Jackson, MS
3350	Houma-Thibodoux, LA	3880	Lafayette, LA
3361	Brazoria	3360	Houston-Brazoria,TX
3500	Iowa City, IA	2120	Des Moines, IA
3580	Jackson, TN	3660	Johnson City-Kingsport--Bristol, TN/VA
3610	Jamestown-Dunkirk, NY	2970	Glens Falls, NY
3870	LaCrosse, WI	3800	Kenosha, WI
4080	Laredo, TX	3810	Kileen-Temple, TX
4100	Las Cruces, NM	200	Albuquerque, NM
4482	Orange County	4481	Anaheim-santa Ana-Garden Grove,CA
4940	Merced, CA	4920	Memphis, TN/AR/MS
5040	Midland, TX	5800	Odessa, TX
5190	Monmouth-Ocean, NJ	5460	New Brunswick-Perth Amboy-Sayreville, NJ
5330	Myrtle Beach, SC	1760	Columbia, SC
5340	Naples, FL	5000	Miami-Hialeah, FL
5350	Nashua, NH	4760	Manchester, NH
5480	New Haven-Meriden, CT	5482	New Haven, CT
5520	New London-Norwich, CT/RI	5482	New Haven, CT
5604	Same name match-merging	5600	Same name match-merging
5607	Same name match-merging	5600	Same name match-merging
5720	Norfolk-VA Beach--Newport News, VA	5721	Norfolk-VA Beach, VA
5950	Orange, NY	5660	Newburgh-Middletown, NY
6010	Panama City, FL	6080	Pensacola, FL
6281	Beaver County	6280	Pittsburg,PA
6460	Same name match-merging	5605	Same name match-merging
6482	Pawtuckett-Woonsocket-Attleboro	6480	providence-Fall River-Pawtucket,RI/MA
6580	Punta Gorda, FL	5960	Orlando, FL
6641	Vancouver,WA	6440	Portland-Vancouver,OR/WA
6820	Rochester, MN	6980	St. Cloud, MN
6890	Rocky Mount, NC	6640	Raleigh-Durham, NC
7361	Oakland	7360	San Francisco-Oaklan-Vallejo,CA
7460	San Luis Obispo-Atascad-P Robles, CA	7400	San Jose, CA
7490	Santa Fe, NM	7480	Santa Cruz, CA
8140	Sumter, SC	3160	Greenville-Spartanburg-Anderson SC
9000	Wheeling	6280	Pittsburg,PA
9270	Yolo, CA	8780	Visalia-Tulare-Porterville, CA
9340	Yuba City, CA	8780	Visalia-Tulare-Porterville, CA
9360	Yuma, AZ	8520	Tucson, AZ

Figure E-1 Geographic Center of Labor Market 3160 and 99425045



E.2 Local Linear Projection for the Expected Wage/Earnings Offer for the Parents

In this Appendix we talk briefly on the local linear projection method we use to obtain the location specific wage/earnings offer for the parents.

For each target prediction (defined by labor market, year, age, race, education, done separately by men and women) we use a weighted linear regression defined as following:

$$\ln w_i = \alpha_1[age_i - bage_i] + \alpha_2[education_i - baseeducation_i] + \alpha_3 \begin{bmatrix} M_i = if\ bsex = female \\ F_i = if\ bsex = male \end{bmatrix} +$$

$\begin{matrix} \alpha_4 \begin{bmatrix} B_i \end{bmatrix} \\ \alpha_5 \begin{bmatrix} H_i \end{bmatrix} \end{matrix}$ if base race= White or, $\begin{matrix} \alpha_4 \begin{bmatrix} B_i \end{bmatrix} \\ \alpha_5 \begin{bmatrix} W_i \end{bmatrix} \end{matrix}$ if base race = Hispanic or, $\begin{matrix} \alpha_4 \begin{bmatrix} W_i \end{bmatrix} \\ \alpha_5 \begin{bmatrix} H_i \end{bmatrix} \end{matrix}$ if base race =black.

The estimated intercept is the prediction for the 'target'. The explanatory variables are defined by the difference between the observation from the sample we use (from Census/CPS dataset) and the "target" combination we use. If it is a dummy variable, then the observations having the same dummy value as the target will receive a weight equal to one in the regression.

We slightly change the traditional kernel estimation that has constant bandwidth to each characteristic, to accommodate the fact that each characteristic should have unique contribution to the wage/earnings outcome. We construct separate weights for each observation in each of these regressions. In particular, for a particular target labor market, target age, target race, target education, we define for each observation i the weight that is the product of the weights for how each of i's characteristics differs from the target.

Only men are used to predict male wage moments, and only women are used to predict women's wage moments. Also, for each target labor market we restricted how far any observation i could be located from the center of the target labor market. In particular, we defined a variable outside_distance(T) as the shortest distance from the center of T's labor market (in multiples of 10 miles) such that there would be at least 500,000 combined census/cps observations across all years of the same gender within a radius of length outside_distance(T). In

particular if the distance of the center of i's conspuma to the center of the target labor market's center is greater than outside_distance(T), then the weight for that observation i is set to zero for target T.

BIBLIOGRAPHY

[1] Aitkin, M and N. Longford (1986). “Statistical Modeling Issues in School Effectiveness Studies.” Journal of the Royal Statistical Society.

[2] Betts, Julian R (1995). “Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth”, Review of Economics and Statistics.

[3] Black, D. A., B. J. Noel, et al. (1999). "On-the-job training, establishment size and firm size: Evidence for economies of scale in the production of human capital." Southern Economic Journal 66(1): 82-100.

[4] Borjas, George J. (1987) “Self-selection and the Earnings of Immigrants”, American Economic Review. 531-553

[5] Card, David, and Alan B Krueger (1992). “Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States, Journal of Political Economy 100, NO. 1 1-40

[6] Childs, T Stephen and Charol Shaksshaft (Fall1986). “A Meta Analysis and Research on the Relationship between Educational Expenditures and Student Achievement”, Journal of Education Finance 12, 249-263

[7] Glass Gene V. and M. L. Smith (1979). “Meta Analysis of Research on Class Size and Achievement’, Educational Evaluation and Policy Analysis 1, 2-16

[8] Grunfeld, Yehuda, and Zvi Griliches (1960). “Is Aggregation Necessarily Bad?” Review of Economics and Statistics 62 NO.1 1-13

[9] Liu, H.Y., Thomas Mroz, Wilbert van der Klaauw (Feb 2007). “Maternal Employment, Migration, and Child Development”, working paper

[10] Hanushek, Eric A (March1998). “Conclusions and Controversies about the Effectiveness of School Resources”, FRBNY Economic Policy Review

[11] Hanushek, Eric A, Steven G. Rivkin, Lori L. Taylor (1996). “Aggregation and the Estimated Effects of School Resources”, NBER Working Paper.

[12] Hanushek, Eric A (1974). “Efficient Estimators for Regressing Regression Coefficients”, the American Statistician 28, NO.2 , 66-69

[13] Heckman, J., A.LayneFarrar, Layne-Farrar, Anne, Todd, Petra. (1996). "Human capital pricing equations with an application to estimating the effect of schooling quality on earnings." Review of Economics and Statistics 78(4): 562-610.

[14] Heckman, J. J. and J. R. Walker (1990). "The Relationship between Wages and Income and the Timing and Spacing of Births - Evidence from Swedish Longitudinal Data." Econometrica 58(6): 1411-1441.

[15] Johnson, George E., and Frank P. Stafford (Spring 1973). "Social Returns to Quantity and Quality of Schooling", Journal of Human Resources 8,139-155

[16] Kain, J. F. and J. M. Quigley (1970). "Measuring Value of Housing Quality." Journal of the American Statistical Association 65(330): 532-&.

[17] Loeb, S. and J. Bound (1996). "The effect of measured school inputs on academic achievement: Evidence from the 1920s, 1930s and 1940s birth cohorts." Review of Economics and Statistics 78(4): 653-664.

[18] Link, Charles R. and Edward Ratledge (1975). "Social Returns to Quantity and quality of Schooling", Journal of Human Resources 8, 139-154

[19] Mcfadden, D. (1976). "Quantal Choice Analysis - Survey." Annals of Economic and Social Measurement 5(4): 363-390.

[20] Mroz, T. A. (1987). "The Sensitivity of an Empirical-Model of Married Womens Hours of Work to Economic and Statistical Assumptions." Econometrica 55(4): 765-799.

[21] Mroz, T., Guilkey, D. (1992). "Discrete factor approximations for use in simultaneous equation models with both continuous and discrete endogenous variables". Mimeo, Department of Economics, University of North Carolina. Chapel Hill.

[22] Mroz, T., Weir, D. (1994). "Random parameters and approximations to stochastic dynamic optimization models with an application to age at marriage and life cycle fertility control in France under the ancient regime". Mimeo, University of North Carolina. Chapel Hill.

[23] Mroz, T. A. (1999). "Discrete factor approximations in simultaneous equation models: Estimating the impact of a dummy endogenous variable on a continuous outcome." Journal of Econometrics 92(2): 233-274.

[24] Murnane, Richard J. (1975). "Impact of School Resources on the Learning of Inner City Children", Cambridge, Mass.: Ballinger.

[25] Murnane, Richard J., and Barbara Phillips (1981). "What Do Effective Teachers of Inner-City Children Have in Common?" Social Science Research 10, no. 1 (March): 83-100.

[26] Oates, W. E. (1969). "Demand and Supply of Public Goods - Buchanan,Jm." Journal of Economic Literature 7(1): 103-104.

[27] Peltzman, Sam (April 1993). "The Political Economy of the Decline of American Public Education." Journal of Law and Economics 36: 331-70.

[28] Poterba, James M. (1996). "Individual and Social Responsibility: Child Care, Education, Medical Care, and Long-term Care in America", 277-304. Chicago: University of Chicago Press/NBER

[29] Rizzuto, Ronald and Paul Watchel (1980). "Further evidence on the Returns to School Quality", Journal of Human Resources 15, 240-272

[30] Tiebout ,Charles M. (Oct 1956) “A Pure Story of Local Expenditures”, Journal of Political Economy 64, 416-424

[31] McFadden, D. and K. Train (2000). "Mixed MNL models for discrete response." Journal of Applied Econometrics 15(5): 447-470.

[32] Todd, Petra E. and Kenneth I. Wolpin (2007). “the Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps”, Journal of Human Capital Vol 1 no. 1

[33] Kennan, John and Walker, James (2008). “The Effect of Expected Income on IndividualMigration Decisions”, University of Wisconsin-Madison and NBER Working Paper