

# Accounting for Time-Varying Unobserved Ability Heterogeneity within Education Production Functions

Weili Ding  
Queen's University

Steven F. Lehrer\*  
Queen's University and NBER

July 2008

## Abstract

Traditional strategies to estimate education production functions do not explicitly allow either the productivity or development of unobserved skills to vary as a child ages, which appears inconsistent with a growing body of scientific evidence. Empirically if one wishes to obtain unbiased parameter estimates of observed educational inputs researchers must properly account for these unobserved skills. In this paper, we introduce a strategy to estimate education production functions that allows for time-varying unobserved ability within individuals and a more general decaying pattern of observed inputs. Using experimental data from Project STAR we present evidence that accounting for both features is important in practice. We find that the effects of unobserved ability on cognitive achievement vary significantly not only between grades in the same subject area but also there is substantial heterogeneity in the impacts within grades.

JEL codes: I20, J24, C33 and C81.

---

\*We would like to thank Richard Murnane and seminar participants at Simon Fraser University, 2008 CSWEP/CEMENT workshop, 2008 CEA Annual meetings and the 2004 AEA annual meetings session on Education for the Disadvantaged and Queen's University for helpful comments and suggestions. We are grateful to Alan Krueger for generously providing a subset of the data used in the study. This paper is a revised version of the second chapter of Weili Ding's University of Pittsburgh 2002 thesis. Lehrer wishes to thank SSHRC for research support. We are responsible for all errors.

# 1 Introduction

Since the landmark publication of the 1966 U. S. Department of Education study titled Equality of Educational Opportunity (aka The Coleman Report), thousands of studies in the economics and education literatures have estimated education production functions to examine whether educational “inputs” correlate with cognitive achievement. Perhaps the major obstacle in production function estimation is that the input decisions that a parent makes depend on their child’s characteristics. Because many of the child’s characteristics that affect these investment decisions are unobserved to the analyst, this gives rise to an endogeneity problem. Intuitively, if a parent adjusts to a change in unobserved innate characteristics by increasing or decreasing their investments depending on whether the change is favorable or not, then these unobserved characteristics and inputs are correlated and biased estimates result. Many researchers interpret these unobserved factors to be either innate ability or unobserved ability heterogeneity.

Economists often use imperfect proxies for unobserved ability or assume their impacts are constant over time or between siblings including twins. This allows the researcher to either (partially) control for this factor or difference it out in the analysis. However, a large and growing multi-disciplinary literature summarized within Knudsen, Heckman, Cameron, and Shonkoff (2006) and Cunha, Heckman, Lochner and Masterov (2006) has demonstrated the malleability of cognitive (and non-cognitive) ability during childhood.<sup>1</sup> These skills are not fixed following conception but rather are related to development of specific brain structures which emerge from both epigenetic and genetic processes. Since unobserved ability heterogeneity is potentially an important contributor

---

<sup>1</sup>Evidence that gaps in unobserved (cognitive) ability between individuals develop at early ages has been documented within economics (Carneiro and Heckman (2003)) as well as the child development literature (e. g. Shonkoff and Phillips (2000)).

to the development of cognitive achievement,<sup>2</sup> it would be advantageous to account for its impacts when estimating education production functions in a more flexible manner than existing methods. More generally, if one wishes to obtain unbiased parameter estimates of educational inputs then researchers must properly control for unobserved ability when estimating education production functions.<sup>3</sup>

This paper considers estimating education production functions that account for time varying effects of unobserved ability heterogeneity, both with and in the absence of ideal data. We introduce a simple empirical approach that permits estimation of the time varying effects of unobserved ability heterogeneity within the standard framework of education production functions.<sup>4</sup> Our empirical

---

<sup>2</sup>Within the labor economics literature the empirical importance of unobserved ability heterogeneity to lifetime welfare has been clearly demonstrated. Keane and Wolpin (1997) report that age 16 measures of unobserved ability endowments account for 90% of the total variance in lifetime earnings. Murnane, Willett, and Levy (1995) find that a substantial fraction of the rise in the returns to education between 1978 and 1986 for young workers is attributable to a rise in the return to ability. Heckman and Vytlačil (2001) find this result robust only for a portion of the sample with high scores (in the fourth quartile) on the Armed Services Vocational Aptitude Battery achievement test.

<sup>3</sup>The idea that similar inputs affect the development of ability and test scores and there are feedbacks has been documented empirically. For example, using scores from the Armed Services Vocational Aptitude Battery achievement test to proxy for unobserved ability, Hansen, Heckman and Mullen (2004) present evidence that ability measured at early ages influences the level of education completed (i.e. high school dropouts, high school graduates, etc.) heterogeneously which affects measures of ability taken at later ages. In this study, we do not jointly estimate a unobserved ability production function using methods such as dynamic factor analysis (e.g. Cunha (2006), Cunha and Heckman (2006) and Cunha, Heckman, and Schennach (2006)) since one must ex-ante assume they know the domains of these skills otherwise they risk introducing measurement error when estimating education production functions. As we detail in the next section, our approach requires the use of instrumental variables and reduces problems of measurement error that arise from using proxies for unobserved ability heterogeneity, which is fundamentally a variable that cannot be measured (yet).

<sup>4</sup>The relationship between empirical specifications of education production functions and the underlying theory is examined in Todd and Wolpin (2003), Boardman and Murnane (1979) and Hanushek (1979). Researchers have also studied the appropriateness of different specifications of an education production function by considering the

strategy allows for very general patterns of the impacts of both observed and unobserved inputs to the education production process.<sup>5</sup> The approach requires at least two years of data on education outputs and inputs and that the unobserved ability heterogeneity enters the education production function in an additively separable manner from the other inputs. Provided there does not exist higher order correlations in the residuals of the education production function, lagged dependent variables can be employed to identify the time-varying impacts of unobserved ability heterogeneity using a GMM procedure. It is important to state explicitly that our empirical approach does not require measures to proxy for unobserved ability and can accommodate heterogeneous impacts of these time-varying unobservables which affect outcomes differently for different individuals and are most likely correlated with the included explanatory variables. Further, quantile regression instrumental variables estimation of the model presents an opportunity to investigate how these impacts vary across the conditional achievement distribution in both different subject areas and at alternative ages.

To illustrate our empirical strategy we use experimental data from Tennessee’s Student/Teacher Achievement Ratio experiment, known as Project STAR. We make use of the feature that teachers were randomly assigned to classrooms in each year of the experiment to overcome important sources of bias in estimating education production functions (Rothstein (2008)). We empirically 

---

functional form (Figlio (1999)), levels of aggregation (Hanushek, Rivkin and Taylor (1996), relevant control variables (Haveman and Wolfe (1995)) and what constitute the appropriate measures of school output (Card and Krueger (1992)).

<sup>5</sup>As such, it nests several popular approaches used in the empirical literature to estimate education production functions. These approaches place implicit assumptions on how the impacts of both observed and unobserved inputs to the production process vary as a person ages. Recently, Todd and Wolpin (2007) used NLSY79-CS data to investigate the assumptions underlying commonly used achievement production functions (assuming the impact of unobserved ability heterogeneity is time invariant) and found little empirical support for these assumptions. In Section 4, we conduct statistical tests using our empirical strategy to determine which (if any) of these alternative empirical restrictions are supported.

demonstrate that it is important to account for the time varying effects of unobserved individual ability heterogeneity, particularly in reading and word recognition. Second, we present evidence that there is substantial heterogeneity in these impacts within each subject area at every grade level. The heterogeneity in these impacts is particularly large in mathematics as in just two years individuals in the top decile of the conditional achievement distribution receive approximately twice the impact from unobserved ability relative to individuals in the lowest decile. While our empirical implementation of this method is within the economics of education, this approach could be used in other contexts where unobserved unit-specific heterogeneity is believed to both play an important role and may have time-varying impacts. For example, it could be used to estimate whether this source of heterogeneity accounts for much of the gaps that develop between groups, countries or even with other individual outcomes such as health.

This paper is organized as follows. In Section 2, we review the general conceptual model of cognitive achievement and introduce our empirical approach. The empirical approach can either provide estimates of the time varying impacts of a fixed stock of unobserved ability heterogeneity or equivalently allow the level of innate ability to vary over time holding its impact constant.<sup>6</sup> Project STAR experimental data is described in Section 3. The empirical results are presented and discussed in Section 4. In this section, we additionally demonstrate that the sign, magnitude and statistical significance of the impact of educational inputs on measures of academic performance is sensitive to restrictions being imposed on both unobserved ability heterogeneity and the empirical specification of the education production function. A concluding section summarizes our findings and discusses direction for future research.

---

<sup>6</sup>In footnote 8 we demonstrate the observational equivalence between these factors. As the title of paper indicates our focus in illustrating the approach will be to estimate the time varying impacts of a fixed stock of unobserved ability heterogeneity. Alternatively, one could interpret this coefficient as the estimated malleability of unobserved skills assuming constant impacts over time.

## 2 Economic Model

We draw on the human capital production function framework introduced by Ben-Porath (1967), and extended by Leibowitz (1974) to the context of investment in children. The general conceptual model depicts the level of achievement  $A_{iT}$  for a given student  $i$  at a point in time  $T$  to be a function of the full history of family, community, school inputs and own innate abilities. These variables may interact with each other in a nontrivial, unknown way. This general model expresses current achievement over time as

$$A_{iT} = f_T(F_{iT} \dots F_{i0}, C_{iT} \dots C_{i0}, S_{iT} \dots S_{i0}, I_i, \epsilon_{iT}) \quad (1)$$

where  $F_{iT}$  is a vector of individual and family characteristics,  $C_{iT}$  is a vector of community variables,  $S_{iT}$  is a vector of school characteristics,  $I_i$  is a vector of unobserved heterogeneity including such factors as student innate abilities, parental tastes, determination, among others and  $\epsilon_{iT}$  is assumed to be distributed with zero mean and no serial correlation. Empirical researchers estimate education production functions to understand the nature of this dynamic process and how specific inputs influence the development of  $A_{iT}$ .

### 2.1 Empirical Cumulative Model

Linearizing the achievement relationship (equation (1)) yields

$$A_{iT} = \beta_{0T} + \beta_{1T}F_{iT} + \beta_{2T}C_{iT} + \beta_{3T}S_{iT} + \beta_{IT}I_i + \left( \sum_{t=0}^{T-1} \beta_{0t} + \beta_{1t}F_{it} + \beta_{2t}C_{it} + \beta_{3t}S_{it} \right) + \epsilon_{iT} \quad (2)$$

Since the regressors can include higher order terms and interaction terms to capture nonlinear relationships, the linearization of the theoretical model imposes few restrictions other than additive

separability of the error term onto the theory.<sup>7</sup> For simplicity, we re-express the relationship as

$$A_{iT} = \beta_T X_{iT} + \beta_{IT} I_i + \left( \sum_{t=0}^{T-1} \beta_t X_{it} \right) + \epsilon_{iT} \quad (3)$$

where  $X_{it}$  is a matrix containing the intercept and all the inputs ( $[1, F_{it}, C_{it}, S_{it}] \forall t$ ). Similarly the relationship in the previous period can be expressed as

$$A_{iT-1} = \alpha_{T-1} X_{iT-1} + \alpha_{IT-1} I_i + \left( \sum_{t=0}^{T-2} \alpha_t X_{it} \right) + \epsilon_{iT-1} \quad (4)$$

Notice the difference in coefficient vectors between equations (3) and (4) as we allow the effects of inputs observed by the analyst on achievement levels to vary freely over time. We do not impose any restrictions on how these impacts vary over time across the full set of education inputs. Reexpressing the relationship in equation (4) as a function of unobserved heterogeneity yields:

$$I_i = \frac{1}{\alpha_{IT-1}} \left( A_{iT-1} - \epsilon_{iT-1} - \sum_{t=0}^{T-1} \alpha_t X_{it} \right) \quad (5)$$

Substituting equation (5) in to equation (3) yields

$$A_{iT} = \beta_T X_{iT} + \frac{\beta_{IT}}{\alpha_{IT-1}} A_{iT-1} + \sum_{t=0}^{T-1} \left( \beta_t - \frac{\beta_{IT}}{\alpha_{IT-1}} \alpha_t \right) X_{it} + v_{iT} \quad (6)$$

where  $v_{iT} = \epsilon_{iT} - \frac{\beta_{IT}}{\alpha_{IT-1}} \epsilon_{iT-1}$ .<sup>8</sup>

---

<sup>7</sup>Variants of this model are also the starting point for analyses of coefficient biases from estimates of education production function. Boardman and Murnane (1979) similarly begin with equation (2) assuming the residual is serially uncorrelated. Hanushek (1979) does not include residuals in equation (2). Todd and Wolpin (2003) include current random shocks in the achievement function given in equation (2) but assume that shocks are serially correlated. The degree of serial correlation will affect how far lagged a dependent variable must be to serve as an instrument. The validity of assumption on the extent of serial correlation could be tested by the Arellano-Bond (1991) test for autocorrelation, which is simple to apply to linear GMM regressions.

<sup>8</sup>Equation (6) is observationally equivalent to an equation where the level of innate ability varies over time  $I_{iT} = \theta_t I_{iT-1}$  but its impact ( $\beta_{IT} = \alpha_{IT-1}$ ) is held constant. Thus, the parameter  $\frac{\beta_{IT}}{\alpha_{IT-1}}$  in equation (6) is observationally equivalent to  $\theta_t$ . The parameter  $\theta_t$  can be interpreted as a growth (or decay) parameter as it explains how the stock of innate ability varies over time.

Direct OLS estimation of equation (6) may not be possible since  $A_{iT-1}$  is correlated with the error term  $v_{iT}$ , which contains  $\epsilon_{iT-1}$ — a component of  $A_{iT-1}$ . Instrumental variables approaches can be used to overcome this endogeneity problem and provide consistent estimates of the parameter  $\frac{\beta_{IT}}{\alpha_{IT-1}}$ , the ratio of the cumulative effect of individual unobserved heterogeneity (i.e. innate ability) between  $T$  and  $T - 1$ . Candidates for instruments in this setting could be suitably lagged endogenous and predetermined variables such as test scores from period  $T-2$  and earlier.<sup>9</sup> Efficient GMM estimation will typically exploit a larger number of instruments at each grade level as more information becomes available. This strategy provides a complete picture of how both observed inputs and unobserved heterogeneity affect achievement levels at different points in time.

Hausman tests comparing instrumental variable and OLS estimates of equation (6) can be used to test for the endogeneity of educational inputs and unobserved heterogeneity. Efficiency gains in this setting are possible by a GLS procedure when the researcher properly accounts for the error component structure. Researchers could also use estimates of equation (6) to conduct specification tests on  $\frac{\beta_{IT}}{\alpha_{IT-1}}$ . Tests on the ratio of this parameter could be used to examine the validity of assumptions on the impacts of unobserved ability heterogeneity that several popular empirical methods adopt to estimate education production functions.

Appendix 1 reviews the three most popular empirical approaches to estimate education production functions. Many of these alternative approaches to estimate education production functions are taken only due to data limitations but also ease in implementation. These approaches are commonly known in the literature as the contemporaneous model, linear growth (or gains) model and value added model. Each approach either implicitly assumes that the impacts of unobserved ability heterogeneity is fixed as a student ages or that the impact does not exist. Yet, even without data on the full history of inputs one can still account for time-varying impacts of unobserved ability

---

<sup>9</sup>Similar to the dynamic panel data literature (Arellano and Bond (1991)), identification of the model via lagged dependent variables as instruments requires restrictions on the serial correlation properties of the error term.

heterogeneity. For example, with only  $m$  periods of data on inputs ( $T - m..T$ ), we can use the same logic that generated equation (6) and express  $A_{iT}$  as

$$A_{iT} = \beta_T X_{iT} + \frac{\beta_I}{\alpha_{IT-1}} A_{iT-1} + \sum_{t=T-m}^{T-1} \left( \beta_t - \frac{\beta_I}{\alpha_{IT-1}} \alpha_t \right) X_{it} + v_{iT}^m \quad (7)$$

where  $v_{iT}^m = \varepsilon_{iT} + \sum_{t=0}^{T-m} \left( \beta_t - \frac{\beta_I}{\alpha_{IT-1}} \alpha_t \right) X_{it} - \frac{\beta_I}{\alpha_{IT-1}} \varepsilon_{iT-1}$ .<sup>10</sup> We can reexpress  $v_{iT}^m$  in terms of  $v_{iT}$  as  $v_{iT}^m = v_{iT} + \sum_{t=0}^{T-m} \left( \beta_t - \frac{\beta_I}{\alpha_{IT-1}} \alpha_t \right) X_{it}$ . Estimation of equation (7) could also be undertaken via instrumental variables or linear GMM estimation.

However, additional difficulties may arise in choosing lagged dependent variables as instruments for  $A_{iT-1}$  since  $v_{iT}^m$  now implicitly contains inputs from earlier periods. Lagged dependent variables are valid instruments provided that  $\exists$  for some period  $l$  s.t.  $(T - m < l < T - 1)$ , s.t.  $\beta_t = \frac{\beta_{IT}}{\alpha_{IT-1}} \alpha_t$  holds  $\forall t = 0..l$ . This is an assumption similar to those underlying various value-added models but has 1) the advantage of allowing for the time varying impacts of unobserved ability heterogeneity, and 2) is less restrictive in assuming that  $A_{it}$  is a sufficient statistic for  $l < (T - l)$  periods of lagged inputs in estimating the achievement equation as opposed to assuming  $A_{iT-1}$  as a sufficient statistic for all  $T - 1$  periods of lagged inputs, thus not imposing any restrictions on the coefficients of the impacts of  $T - l$  to  $T$  period inputs on current achievement. Of course, if one has access to an exogenous variable other than lagged dependent variable, then this assumption on how past inputs enter into the current education production process is unnecessary.

### 3 Data

We use data from Tennessee's highly influential class size experiment, Project STAR, to illustrate our empirical strategy. This experiment was conducted for a cohort of students in 79 schools over a four-year period from kindergarten through grade 3. Within each participating school, incoming

---

<sup>10</sup>Note if  $l > 1$ ,  $A_{iT} = \beta_T X_{iT} + \frac{\beta_I}{\alpha_{IT-1}} A_{iT-1} + \sum_{t=T-l}^{T-1} \beta_t X_{it} + \sum_{t=T-m}^{T-1} \left( \beta_t - \frac{\beta_I}{\alpha_{IT-1}} \alpha_t \right) X_{it} + v_{iT}^m$

kindergarten students were randomly assigned to one of the three intervention groups: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher’s aide). Additionally in each year of the experiment teachers were also randomly assigned to the classrooms in which they would teach.

While our aim is not to contrast structural parameter estimates with the experimental estimate, this dataset has four features which make it ideal to illustrate the empirical strategy proposed above. First, strictly speaking one would need input data from a child’s conception to estimate education production functions. Randomization ensures that the requirement of exogeneity for the inputs holds in the initial period of analysis. Omitting pre-kindergarten inputs should not affect the coefficient estimate on class size or the other structural parameters in kindergarten.<sup>11</sup> Second, random assignment overcomes selection bias that arises not solely by decisions made by parents themselves but also by school principals. School inputs are well known to be choice variables and with non-experimental data we would be required to find credible sources of exogenous variation to identify their impacts. Further, since teachers were re-randomized to classrooms each year we can obtain unbiased estimates of the effects of both current and past teacher characteristics.<sup>12</sup> Third, this data set reduces measurement error from aggregation bias by precisely matching each student to the classroom and teacher within a school so that we can focus on estimates of the time-varying impacts of unobserved ability. Finally, Project STAR was conducted for children between Kindergarten to grade 3, stages in the lifecycle child development specialists have suggested that either the impact or stock of cognitive ability is malleable.

At the end of each school year the majority of the students completed multiple exams to measure

---

<sup>11</sup>The standard error or the precision of the estimates may be affected.

<sup>12</sup>Rothstein (2008) presents evidence from North Carolina that teacher assignments to students are non-random. While the classroom assignment process is the responsibility of school principals, Jacob and Lefgren (2007) present evidence that parents often have strong preferences for specific teachers and are willing to advocate for them which further influences class assignment.

their performance in different dimensions. In this paper, our outcome measures ( $A_{iT}$ ) are total scaled scores from the Reading, Mathematics, Word Recognition sections of the Stanford Achievement test.<sup>13</sup> Scaled scores are calculated from the actual number of items correct adjusting for the difficulty level of the question to a single scoring system across all grades. Scaled scores are usually not comparable across different tests, but within the same test they have the advantage that a one point change on one part of the scale is equivalent to a one point change on another part of the scale. This offers an important advantage in the identification of the ratio of the effects of unobserved heterogeneity in between two periods  $\frac{\beta_{IT}}{\alpha_{IT-1}}$ . If the achievement measures in alternative years are not measured in units from the same scale, for example SAT scores and GRE scores, estimates of  $\frac{\beta_{IT}}{\alpha_{IT-1}}$  will combine information on the ratio of the effects of unobserved heterogeneity with the ratio that places these scores on a similar metric.

A challenge with using data from Project STAR exists since violations to the experimental protocol were prevalent. By grade three over 50% of the subjects who participated in kindergarten left the STAR sample and approximately 10% of the remaining subjects switch class type annually. Additionally, Ding and Lehrer (2008) present evidence of selective attrition and demonstrate that the conditional random assignment of the newly entering students failed in the second year of the experiment as among this group of students those on free lunch were significantly more likely to be assigned to regular (larger) classes.<sup>14</sup> In order to minimize issues related to changing composition of the sample affecting the estimates of the time varying impacts of unobserved ability heterogeneity, we only include students who participated in all four years of Project STAR and completed exams

---

<sup>13</sup>The Stanford Achievement Test is a norm-referenced multiple-choice test designed to measure how well a student performs in relation to a particular group, such as a representative sample of students from across the nation. Norm-referenced tests are commercially published and are based on skills specified in a variety of curriculum materials used throughout the country. They are not specifically referenced to the Tennessee curriculum.

<sup>14</sup>It should also be noted that attendance of kindergarten was not mandatory in Tennessee and students who entered school in grade 1 may differ in unobservables to those started in kindergarten.

in all three subject areas each year in this study.<sup>15</sup>

Summary statistics for this sample are provided in Table 1. Each column presents summary information on this cohort of students with complete data at different grade levels. The percentage of this sample that receives small class treatment increases by almost one third over this four year period. While there are few differences in the percentage of the sample on free lunch across the grades, there are numerous transitions between grade levels. Each year approximately 15% of the students change their free-lunch status. Since our test scores are scaled scores they increase across the grades. Not surprisingly, there is a increase in the variance of both reading and word recognition tests scores over this period. However, there is reduced dispersion in math test scores. Teachers in higher grades on average have more years of experience. In all of our empirical specifications the matrix  $X_{it}$  consists of class size, school effects, years of teaching experience, the education level and race of the teacher, the gender, race and free lunch status of the student  $i$  in year  $t$ .<sup>16</sup>

## 4 Results

### 4.1 How Should We Treat Unobserved Ability?

In this section, we present evidence that accounting for time-varying impact of unobserved ability heterogeneity is important. Table 2 presents instrumental variable estimates of the ratio of the

---

<sup>15</sup>Note the results are also robust to using the full sample of kindergarten students where the samples are reweighted by either series logit estimates of the probability of remaining in the sample or the probability of writing the exam the previous academic year.

<sup>16</sup>These variables are exactly the same as those used in the base specifications in Krueger (1999). For robustness we replicated the entire analysis with two alternative specifications that allowed teacher experience to have nonlinear effects. The first approach allowed different impacts in each of the first three years and the second approach included experience up to a cubic. All of the results discussed in the next section are robust to these alternative treatments of teacher experience.

effects of unobserved individual heterogeneity from equation (7) using inputs from kindergarten onwards with two alternative instrument sets. We first use initial class assignment by itself as an instrument since due to random assignment it should be uncorrelated with unobservables to the production process at every grade level.<sup>17</sup> Second, we use two or more periods lagged achievement scores in all subject areas in the earlier grades.<sup>18</sup> Employing only the initial random assignment to class type as an instrument provides imprecise and statistically insignificant estimates of  $\frac{\beta_{IT}}{\alpha_{IT-1}}$  in each subject and grade. The sign and magnitude of  $\frac{\beta_{IT}}{\alpha_{IT-1}}$  varies substantially with this instrument. Examination of the first stage regression (Appendix Table 1) demonstrates that this is a very weak instrument by conventional criteria.

Instrumental variable estimates with two or more periods of lagged achievement scores as instruments provide more precise evidence on  $\frac{\beta_{IT}}{\alpha_{IT-1}}$ . The estimated impact of unobserved ability heterogeneity appears fairly constant across grades in mathematics, where the contribution of unobserved ability declines slightly (approximately 11%) between grades two and three. Yet in both grades, the constraint that  $\frac{\beta_{IT}}{\alpha_{IT-1}} = 1$  is supported. This constraint is firmly rejected in both grades 2 and 3 reading and the grade 3 word recognition test. Similar to mathematics, on average the estimated impacts of unobserved ability heterogeneity on test scores in both reading and word recognitions decline. The estimated magnitude of the time varying effect declines (on average) by approximately 32% between grades one to three in reading and 16% between grades two and three

---

<sup>17</sup>Krueger (1999) verified whether individuals attended the class type to which they were assigned for 18 of the 79 STAR schools. 99.7% of the kindergarten students attended the class type to which they were assigned. However, if kindergarten class type is being used to instrument later class size and kindergarten class size is omitted from the estimating equation, class type may not be a valid instrument based on the cumulative model of achievement.

<sup>18</sup>The second instrument set could expand in higher grades as more past test scores are available to serve as additional instruments presenting efficiency gains. That is, when estimating the grade 2 achievement equations, test scores from kindergarten can be used as instruments but both kindergarten and grade 1 test scores could be instruments for the grade 3 achievement equation.

in word recognition.<sup>19</sup> Finally, the IV estimates in Table 2 reject the Null hypothesis  $\beta_{IT} = 0$  in all subject areas in grades 2 and 3.

Intuitively, these empirical results conform with ideas from the education literature that students require more cumulative knowledge (i.e. literacy) in reading and word recognition, thus less reliant on unobserved “ability”. While mathematics knowledge acquisition is also a gradual process the structure of test questions changes sharply as a child ages. Mathematics tests in grades two and three focus less on recognizing shapes and numbers and more on problem solving which requires the development of new mental skills to visualize problems (as opposed to sounds or shapes). Taken together, the results in Table 2 suggest that one must account for unobserved heterogeneity in a flexible manner both across time and different subjects.

In order to examine the importance of accounting for unobserved ability heterogeneity in education production functions we calculated the partial R-squared for this variable. The partial R-squared ranged between 20 to 40% of the variation in test scores. The values were close to 40% in both grades 2 and 3 reading and mathematics. In all subject areas and grades the inclusion of unobserved ability heterogeneity accounted for more than twice of the variation in test scores outcomes relative to that which is explained by the full set of current and past observed education inputs. This analysis illustrates the empirical importance of unobserved ability heterogeneity in the production of different measures of cognitive achievement.

Hausman tests between OLS and IV estimates of equation (6) reject both the Null of exogeneity for the entire coefficient vector in all subject areas as well as  $\frac{\beta_{IT}}{\alpha_{IT-1}}$  by itself in grades 2 and 3. Thus, to estimate education production functions with Project STAR one should account for individual specific unobserved heterogeneity that is correlated with the full set of inputs.

To assess the suitability of our instruments we consider a variety of specification tests. Appendix Table 1 indicates that weak instruments do not appear to be a concern for the instrument set

---

<sup>19</sup>Note the results are robust to using a single two period lagged test score in the same subject area as an instrument.

containing lagged test scores. The first stage F-statistics of the hypothesis that the coefficients on the excluded instruments are zero range from 257.95 to 464.25, with a p-value of 0.00 in all cases. Additionally, the individual coefficient on most of the instruments is significant at the 1% level. Not only is there a strong first stage relationship but J tests provide little evidence against the overidentifying restrictions. In contrast and as noted earlier in this section, an F-test for the joint significance of the initial treatment assignment indicator in the first stage demonstrates that this is a very weak instrument using conventional criteria. Finally, Arellano-Bond (1991) test for autocorrelation reject that there is second order serial correlation in the residuals, increasing our confidence that the statistical properties of the instruments are met.

We next estimate equation (6) via the instrumental variables for quantile regression (QRIV) estimator introduced in Cheronukov and Hansen (2005) to determine if there are different impacts from unobserved ability in different parts of the achievement distribution within grades. In other words, this econometric estimator allows us to examine whether there are heterogeneous quantile treatment effects over the entire population.

Figure 1 presents QRIV estimates of  $\frac{\beta_{IT}}{\alpha_{IT-1}}$  and its 95% confidence interval as well as the standard instrumental variables point estimate for each subject in grades 2 and 3. Notice that in all subject areas, there is clear evidence of substantial heterogeneity in  $\frac{\beta_{IT}}{\alpha_{IT-1}}$ . Statistical tests within grades reject the assumption of constant effect across the conditional achievement distribution in all subject areas. At many quantiles in most of the panels contained in Figure 1, the linear IV estimate is not contained within the 95% confidence interval of the QRIV estimate. With the exception of grade two word recognition individuals at higher deciles generally experience larger impacts from unobserved ability heterogeneity.<sup>20</sup> For instance, in mathematics the impact at higher decile is

---

<sup>20</sup>Note since the specifications include a large number of explanatory variables caution should be taken with estimates at the extreme quantile (5/95) as the asymptotics rely on there being enough observations on both sides of the quantile in order to apply a conditional central limit theorem. More details and rules of thumbs are provided in Chernozhukov (2000). The full set of QRIV estimates is available from the authors by request.

statistically greater than 1, whereas the impact in the lowest deciles is significantly below 1. The impact at the highest deciles is approximately 54% larger in magnitude relative to the impact in the lowest deciles. The gap between an individual at the highest quantile in both grades 2 and 3 relative to an individual at the lowest quantile is over 115%.

Figure 1 suggests that large differences in the impacts of unobserved ability heterogeneity across the population already emerged at early ages. In each grade, the gaps in the impacts from unobserved ability across deciles are largest in mathematics and substantially smaller in word recognition. Across grades, the gaps between the highest and lowest quantile are fairly constant in mathematics but decrease by a large fraction in both reading and word recognition. While Table 1 reported that on average the impact of unobserved ability was not significantly different from one in mathematics, Figure 1 presented substantial heterogeneity in these estimated impacts across the distribution. This heterogeneity further demonstrates that traditional differencing approaches of education production functions may not be appropriate for many individuals since at several quantiles unobserved ability does not evolve at a constant unitary rate ( $\frac{\beta_{IT}}{\alpha_{IT-1}} = 1$ ). In addition, from a policy perspective estimating quantile impacts of inputs to an education production function (in addition to mean impacts) is likely of importance since social costs associated with poor human capital development exist primarily at the low end of the achievement distribution, with the costs increasing substantially at the very low end.

## 4.2 Comparing Alternative Empirical Approaches

Estimating equations (6) and (7) using lagged dependent variables as instruments not only allows researchers to recover the time varying impacts of unobserved ability but also is more flexible in the restriction that the method imposes on how the impact of past observed education inputs decay relative to the popular approaches to estimate education production functions. This section demonstrates that these differences are indeed important in practice as the less flexible methods

reviewed in Appendix 1 would lead to substantially different estimates regarding the effectiveness of education inputs compared to that obtained from equation (7). The empirical methods discussed in Appendix 1 include current education inputs as explanatory variables and are known as i) the contemporaneous model, which assumes full and complete decay of the effects all past observed and unobserved inputs  $\beta_t = 0 \forall t \in [0..T - 1]$  and  $\beta_{IT} = 0$ , ii) the linear growth model uses gains in test scores as a dependent variable, assumes that the effects all past observed and unobserved inputs do not decay,  $\beta_t = \beta \forall t$ ,  $\beta_{IT} = \alpha_{IT}$ , and iii) value added model additionally includes  $A_{iT-1}$  as an explanatory variable, assuming that  $A_{iT-1}$  is a sufficient statistic for all past observed and unobserved inputs.<sup>21</sup>

Table 3 illustrates how the sign, magnitude and statistical significance of the estimated current class size coefficient from equation (7) using the set of lagged achievement measures as instruments differ substantially from those obtained by using the popular approaches. The first column contain instrumental variable estimates of equation (7) where the lagged achievement scores identify  $\frac{\beta_{IT}}{\alpha_{IT-1}}$  and is used as a benchmark since it imposes fewer restrictions on the effects of both observed and unobserved inputs to the production process. The benchmark estimates suggest that current class size is negatively and significantly related (at the 10% level) to achievement on grade 2 word recognition, grade 3 reading and word recognition exams. There does not exist any evidence indicating a significant relationship between current class size and achievement in mathematics.

Columns 2, 3 and 4 of Table 3 present estimates of the effects of current class size from estimates of the value added, linear growth and contemporaneous models respectively. Placing restrictions

---

<sup>21</sup>We use identical terminology to Todd and Wolpin (2007) to describe these empirical models. Within the economics of education literature other names do exist. It should also be noted that all of the empirical methods described in Appendix 1 can be nested within equations (7), presenting an opportunity to conduct a variety of specification tests. We conducted a number of specification tests and found that in all grades and subject areas we are able to reject the restrictions that underlie each of the three empirical approaches described in Appendix 1. Further, the results with data from grade 3 suggest that while  $A_{iT-1}$  is not a good sufficient statistic  $A_{iT-2}$  indeed shows promise.

on the impacts of observed and unobserved inputs to the education production function leads to substantially different estimates and policy recommendations. Consider estimates of the impact of class size on achievement from the linear growth model which uses changes in test scores as a dependent variable (equation (10)) presented in column 2. A reader would reach dramatically different conclusions if this model were estimated in place of the more general approach presented in the column 1 of Table 3. The estimated impact of class size on achievement is of the opposite sign to that presented in column 1 in all subject areas in grade 2 and grade 3 mathematics.

Turning to estimates from the value added model (equation (9)) presented in column 3 which includes lagged achievement as an independent exogenous explanatory variable but only current educational inputs in the specification, a reader would now conclude that class size does not significantly affect performance on the grade 2 word recognition exam. In addition, the magnitude of the estimated impacts of current class size on both the grade 3 reading and word recognitions tests from the value added model are approximately 50% in magnitude compared to column 1.

Estimates of the contemporaneous model (equation 8) are presented in the column 4 of Table 3 and present an extremely different picture regarding the effectiveness of class size. This model assumes that past inputs decay completely and only current inputs affect academic achievement. A reader would conclude that in all subject areas there is a large statistically significant benefit from reduced class sizes. The magnitude of these benefits peaks in grade one and then decreases in higher grades. Whereas estimates from the contemporaneous model would present strong evidence in favor of class size reductions, none of the other approaches presented in Table 4 would provide empirical support for large class size reductions as the structural parameter is not significantly negative in sign in all subject areas in grades 2 and 3.

While the results presented in columns 2, 3 and 4 differ substantially from that presented in column 1 of Table 3 a concern could be that this is simply due to the manner in which we control for lagged educational inputs and not the control for accounting for time varying unobserved ability

heterogeneity that is correlated with the inputs. Ignoring past observables (which equivalently places strong restrictions on how lagged inputs decay) also has large impacts on the coefficients on the free lunch variables which result in estimates that overstate the importance of this factor on current achievement. In columns 5, 6 and 7 we consider specifications of the education production function that control for the full history of observed education inputs but i) ignore the role of unobserved ability, ii) allow for the role of unobserved heterogeneity but assume that it is uncorrelated with educational inputs, and iii) allow for the role of unobserved heterogeneity but assume that it has a constant effect across successive grades and is correlated with educational inputs.<sup>22</sup> If we ignore unobserved ability heterogeneity or treat it as being strictly exogenous to the other inputs different patterns emerge on the effect of class size on current achievement. This is illustrated in column 5 and 6 of Table 3 which report that current class sizes is ineffective on the grade three reading and word recognition tests. The GLS estimates in column 6 are substantially more precise than the OLS estimates in column 5. Further, the GLS estimates are approximately 25% smaller in magnitude in grade one. Estimates in column 7 assume that unobserved ability has a constant effect exhibit minor changes on the coefficients of contemporaneous class size presented in column 1. The limited differences is due in part to the fact that specification tests only reject the assumption that  $\frac{\beta_{IT}}{\alpha_{IT-1}} = 1$  in three of the six tests in grades 2 and 3. Finally and as discussed in the preceding section Hausman tests suggests that we must control for the endogeneity of this term.

Estimates of the impact of current class size on achievement barely exhibit any differences between columns 1 and 8 of the top panel of Table 3. These columns differed in the amount of lagged years of observed inputs included in the specification of equation (7). While specification tests on the further lagged inputs contained in the specifications in column 1 suggest they should

---

<sup>22</sup>Specifically in column 5 and 6 we estimate equation  $A_{ijT} = \beta_T X_{ijT} + \beta_I I_i + (\sum_{t=k}^{T-1} \beta_t X_{ijt} + \rho_t \epsilon_{ijt}) + \epsilon_{ijT}$ . by OLS and GLS respectively where k stands for kindergarten. Column 7 present estimates of equation 6 where  $\frac{\beta_{IT}}{\alpha_{IT-1}}$  is restricted to equal 1. It is possible to recover the structural parameters of lagged inputs using the sequential difference procedure described in Ding and Lehrer (2008).

be included, their exclusion did not dramatically affect the estimates on current class size.

The results of Table 3 clearly demonstrates that the different empirical approaches present substantially different pictures of the effectiveness of smaller classes, readers must consider the sensitivity of any findings to the credibility of the assumptions that the alternative approaches implicitly impose on the education production process when interpreting the evidence.<sup>23</sup> Since estimates of equation (7) include lagged educational inputs one could also notice that the manner in which home and school inputs decay varies in an unsystematic manner. This is suggestive that the restrictions imposed on both observed and unobserved inputs by traditional strategies to estimate education production functions could be excessively restrictive and may additionally bias parameter estimates of observed educational inputs.

Since estimates of the impact of current class size on achievement barely exhibit any differences between columns 1 and 8 of the top panel of Table 3, we next examined if using fewer years of data on lagged inputs affected the sign, significance and magnitude of any of the other current education inputs. Table 4 compare parameter estimates of contemporaneous home and school inputs from equation (7) using either the full STAR data or only one period of lagged inputs where the set of lagged achievement measures are used as instruments to identify  $\frac{\beta_{IT}}{\alpha_{IT-1}}$ . Notice that the coefficients on the ratio of the impact of unobserved ability heterogeneity as well as the school inputs including class size are virtually identical in sign, magnitude and statistical significance in all subject areas between these inputs.<sup>24</sup> Our results using Project STAR data suggest the data requirements to estimate equation (7) may not be extensive in practice. To summarize, the empirical approach introduced in this paper is more flexible than those described in Appendix 1 and as there are an increasing number of rich longitudinal education datasets being collected around the world,

---

<sup>23</sup>Similarly estimates of causal impacts from Project STAR differ based on the assumptions researchers use to handle violations to the experimental protocol (e. g. Krueger (1999) compared with Ding and Lehrer (2008)).

<sup>24</sup>Only the sign but not statistical significance on the impact of current free lunch status on reading achievement in both grades two and three differs between these columns.

we encourage researchers to consider adopting more general estimation strategies that place fewer restrictions on the underlying model in order to obtain unbiased parameter estimates that have important policy implications.

## 5 Conclusion

In the economics of education literature, when it comes to estimating education production functions researchers often implicitly assume that both the impact and stock of unobserved ability are constant over time. This appears inconsistent with a rapidly growing body of scientific evidence which indicates that the productivity and development of these unobserved skills vary substantially over the lifecycle especially in childhood. In this paper, we introduce an instrumental variables method to estimate education production functions that allows for time-varying unobserved ability within individuals and imposes weaker restrictions on the decaying pattern of inputs. We present evidence that accounting for both features is important empirically. Our results suggest that unobserved ability is correlated with inputs to the production process. We find the impacts of unobserved ability on achievement between grade one to grade three diminish by approximately 32% and 15% in reading and word recognition (on average) respectively. Since the effects of unobserved ability on cognitive achievement vary between three grade levels even in the same subject area, traditional differencing approaches of education production functions such as the within individual transformation may be restrictive. Further, the impacts of unobserved ability vary substantially over the population particularly in mathematics. Thus, even when on average this ability impact appears constant over time such as in math, traditional differencing may still be invalid for individuals at many quantiles of the achievement distribution. We find that specifications of the education production function that place strong implicit restrictions on how lagged inputs decay result in substantial differences in the estimated sign, significance and magnitude of current class size and free lunch variables

on academic performance. Finally, this empirical strategy introduced in this paper may extend beyond education production functions and have implications for empirical researchers that seek to explain gaps between groups or countries as well as those working with other cumulative models of individual development such as health production.

In future research we hope to extend the methodology described in this paper to develop an estimable panel data model in which the individual effect has multiple components and each of these components is time-varying. Developing estimable education production functions from the underlying economic model that could be estimated by recent econometric methods that assume the unobservable individual effects has a factor structure (i.e. Bai (2005) and Ahn et al. (2007)) could shed new policy relevant insights. For instance it could potentially allow us to identify the time varying impacts of both different dimensions of unobserved skills (i.e. cognitive vs. non cognitive) as well as observed inputs on measures of academic performance to reveal which targeted education interventions could yield the largest returns.

## Appendix I: Traditional Methods to Estimate Education Production Functions

The three most popular empirical approaches in the economics of education literature to estimate education production functions impose assumptions on equation (2) regarding how the impacts of observed historical inputs into the production function decay.<sup>25</sup> These approaches additionally assume that (if non-zero) the contemporaneous effects of unobserved heterogeneity are fixed as students age.

The first approach is often referred to as the contemporaneous education production function as it only includes current measures of education inputs as explanatory variables. Researchers estimate

$$A_{iT} = \beta' X_{iT} + \varepsilon_{iT}^c, \quad (8)$$

where  $\varepsilon_{iT}^c = \beta_{IT} I_i + (\sum_{t=0}^{T-1} \beta_t X_{it} + \rho_t \varepsilon_{it}) + \varepsilon_{iT}$ . Unbiased parameter estimates from equation (8) require that past inputs to the production process and unobserved ability decay immediately.<sup>26</sup>

The second approach requires that the researcher has access to two periods of achievement measures and is commonly called a value added model. This model reexpresses the achievement function as:

$$A_{ijT} = \beta_T X_{ijT} + \delta A_{ijT-1} + \varepsilon_{ijT}^L \quad (9)$$

where  $\varepsilon_{ijT}^L = \varepsilon_{iT} + (\beta_{IT} - \delta \alpha_{IT-1}) I_i + \sum_{t=0}^{T-1} (\beta_t - \delta \alpha_t) X_{it} + \sum_{t=0}^{T-1} (\rho_t - \delta \varsigma_t) \varepsilon_{it}$ . The inclusion of  $A_{ijT-1}$  in the regression equation (9) is to pick up a variety of confounding influences including the prior, and often unrecorded as well as unobserved history of parental, school and community effects. Consistent and unbiased parameter estimates from equation (9) require that the effect of

---

<sup>25</sup>We provide a brief review below and guide the reader to Todd and Wolpin (2003) for a more comprehensive discussion.

<sup>26</sup>This requires  $\beta_t = 0 \forall t \in [0..T-1]$  and  $\beta_{IT} = 0$ . Parameter estimates of current inputs would be biased if past inputs or unobserved ability both directly affect current achievement and are correlated with current inputs.

both observed and unobserved factors in the production process to decay over time at the same rate as no past inputs and shocks are left unrepresented by  $A_{it-1}$ .<sup>27</sup>

The third approach is often referred to as either the linear growth or the gains model since the estimating equation is expressed as a function of the growth rate in test scores ( $\Delta A_{iT} = A_{iT} - A_{iT-1}$ ),<sup>28</sup> as

$$\Delta A_{iT} = \beta' X_{iT} + \tilde{\varepsilon}_{iT} \quad (10)$$

where  $\tilde{\varepsilon}_{iT} = \varepsilon_{iT} + (\beta_{IT} - \alpha_{IT-1})I_i + \sum_{t=0}^{T-1}(\beta_t - \alpha_t)X_{it} + \sum_{t=0}^{T-1}(\rho_t - \varsigma_t)\varepsilon_{it}$ . Unbiased and consistent parameter estimates from equation (10) require that past inputs to the production process have constant impacts on achievement at different points in time.<sup>29</sup>

---

<sup>27</sup>This requires  $\beta_t = \delta\alpha_t, \beta_I - \delta\alpha_I$  and that any serial correlation is constant over time. Thus, the empirical strategy assumes  $A_{iT-1}$  to be a sufficient statistic of all the previous influences, which means that  $A_{iT-1}$  is a state variable following a Markov process.

<sup>28</sup>This was introduced in Hanushek (1979), who noted that if one were to assume that unobserved heterogeneity had a constant effect then by differencing equation (4) from equation (3) removes  $I_i$  from the regression equation.

<sup>29</sup>This assumption is fairly restrictive as it implies that having a good second grade math teacher has the same impact on an achievement measure when an individual was in college as when she was a second grader. Note, a variant of the linear growth model allows unobserved heterogeneity to affect the growth rate of achievement. Researchers estimate

$$\Delta A_{iT} = \beta' X_{iT} + \gamma'_I I_i + \tilde{\varepsilon}_{iT} \quad (11)$$

and several of these researchers argue that this would result in less bias for the empirical model than estimating equation (9). For example, Zimmer and Toma (1999 p.80) state “by estimating the value added model the biases are reduced below that which would result from estimating levels of achievement because only the growth effect of innate ability is omitted.” Such claims are unfounded since the focus is misplaced on the empirical model rather than the underlying model of cumulative achievement. Empirically, without data on innate abilities, one can not distinguish between estimates of equation (10) or equation (11).

## References

- [1] Ahn, S. C., Y. H. Lee and P. Schmidt (2007), "Panel Data Models with Multiple Time-Varying Individual Effects," *Journal of Productivity Analysis* 27(1), 1-12.
- [2] Arellano, M. and S. Bond (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies* 58(2), 277-297.
- [3] Bai, J., (2005), "Panel Data Models with Interactive Fixed Effects," *mimeo*, New York University.
- [4] Ben-Porath, Y. (1967), "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy* 75(4), 352-365.
- [5] Boardman, A. E. and R. J. Murnane (1979), "Using Panel Data to Improve Estimates of the Determinants of Educational Attainment," *Sociology of Education* 52(1), 113-121.
- [6] Card, D. and A. 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(1), 1-40.
- [7] Carneiro, P. and J. J. Heckman (2003). Human Capital Policy. In J. J. Heckman, A. B. Krueger, and B. M. Friedman (Eds.), *Inequality in America: What Role for Human Capital Policies?*, Cambridge, MA: MIT Press.
- [8] Coleman, J. S., E. Q. Campbell, C. J. Hobson, J. McPartland, A. M. Mood, F. D. Weinfeld and R. L. York, *Equality of Educational Opportunity*, (Washington D.C.: U.S. Government Printing Office, 1966).
- [9] Chernozhukov, V., and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245-261.
- [10] Chernozhukov, V. (2000) Conditional Extremes and Near-Extremes: Estimation, Inference, and Economic Applications," Ph.D. Dissertation, Stanford University.
- [11] Cunha, F. and J. J. Heckman (2006), "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," forthcoming in *Journal of Human Resources*.
- [12] Cunha, F., J. J. Heckman, and S. M. Schennach (2006), "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *mimeo*, University of Chicago.

- [13] Cunha, F., J. J. Heckman, L. Lochner and D. Masterov (2006), "Interpreting the Evidence on Life Cycle Skill Formation," in E. Hanushek and F. Welch, (eds.), *Handbook of the Economics of Education*, North Holland: Amsterdam
- [14] Ding, W. and S. F. Lehrer (2008), "Estimating Treatment Effects from Contaminated Multi-Period Education Experiments: The Dynamic Impacts of Class Size Reductions," forthcoming in the *Review of Economics and Statistics*.
- [15] Figlio, D. N., (1999), "Functional Form and the Estimated Effects of School Resources," *Economics of Education Review* 18(2), 241-252.
- [16] Hansen, K. T., J. J. Heckman and K. J. Mullen (2004), "The Effect of Schooling and Ability on Achievement Test Scores," *Journal of Econometrics* 121(1-2), 39-98.
- [17] Hanushek, E. A., S. G. Rivkin and L. L. Taylor (1999), "Aggregation and the Estimated Effects of School Resources," *Review of Economics and Statistics* 78(4), 611-627.
- [18] Hanushek, E. A., (1979), "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," *Journal of Human Resources* 14(3), 351-388.
- [19] Heckman, J. J. and E. Vytalacil (2001), "Identifying The Role of Cognitive Ability in Explaining The Level of and Change in The Return to Schooling," *Review of Economics and Statistics* 83(1), 1-12.
- [20] Heckman, J. J., (2000), "Policies to Foster Human Capital," *Research in Economics* 54(1), 3-56.
- [21] Jacob, B. and L. Lefgren (2007), "What Do Parents Value in Education? An Empirical Examination of Parents' Revealed Preferences for Teachers," *Quarterly Journal of Economics* 122(4), 1603-1637.
- [22] Keane, M. P. and K. I. Wolpin (1997), "The Career Decisions of Young Men," *Journal of Political Economy* 105(3), 473-522.
- [23] Krueger, A. B., (1999) "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics* 114(2), 497-532.
- [24] Leibowitz, A., (1974), "Home Investments in Children," *Journal of Political Economy* 82(2), S111-131.
- [25] Murnane, R., J. Willett, and F. Levy (1995), "The Growing Importance of Cognitive Skills in Wage Determination," *Review of Economics and Statistics* 77(2), 251-266.

- [26] Rothstein, J., (2008), “Do Value-Added Models Add Value?,” *mimeo*, Princeton University.
- [27] Shonkoff, J. and D. Phillips eds. (2000), “*From Neurons to Neighborhoods: The Science of Early Childhood Development*,” Washington DC: National Academy Press.
- [28] Todd, P. E. and K. I. Wolpin, (2007), “The Production of Cognitive Achievement in Children: Home, School and Racial Test Score Gaps,” *Journal of Human Capital* 1(1) 91-136.
- [29] Todd, P. E. and K. I. Wolpin (2003), “On the Specification and Estimation of the Production Function for Cognitive Achievement,” *Economic Journal* 113(1), 3-33.
- [30] Zimmer, R. and E. Toma (1999), “Peer Effects in Private and Public Schools Across Countries,” *Journal of Policy Analysis and Management* 19(1), 75-92.

Table 1: Summary Statistics on Sample of Project STAR Participants who Participated in Each Year of the Experiment and Have Completed all Reading, Mathematics and Word Recognition Exams.

	Kindergarten	Grade One	Grade Two	Grade 3
Class Size	19.914 (3.827)	20.334 (4.017)	20.217 (4.118)	20.400 (4.441)
Receiving Small Class Treatment	0.314 (0.464)	0.350 (0.477)	0.373 (0.484)	0.396 (0.489)
Math Test Score	500.038 (44.979)	545.939 (40.405)	594.427 (43.499)	627.977 (40.181)
Reading Test Score	445.673 (31.438)	541.754 (52.412)	599.326 (43.390)	625.634 (37.125)
Word Recognition Test Score	444.702 (37.295)	532.811 (46.788)	600.021 (47.118)	622.771 (43.932)
Free Lunch Status	0.359 (0.480)	0.371 (0.483)	0.354 (0.478)	0.353 (0.478)
Student is White of Asian	0.753 (0.432)	0.753 (0.432)	0.753 (0.432)	0.753 (0.432)
Student is Female	0.518 (0.500)	0.518 (0.500)	0.518 (0.500)	0.518 (0.500)
Teacher Race is Non-White	0.129 (0.335)	0.140 (0.347)	0.178 (0.383)	0.165 (0.372)
Teacher has a Masters Degree	0.377 (0.485)	0.343 (0.475)	0.363 (0.481)	0.443 (0.497)
Teacher Years of Experience	9.447 (5.497)	11.713 (8.625)	13.076 (8.567)	13.547 (8.471)

Note: Each cell reports the mean and standard deviations in parentheses. There are 2239 students who participated and completed all three exams in each year of the experiment.

Table 2: Instrumental Variable Estimates of the Ratio of the Effects of Unobserved Individual Heterogeneity on Achievement at Various Grade Levels by Subject

Subject Area	IV SET 1 Random Class Type Assignment			IV SET 2 Two or More Period of Lagged Test Scores		
	Mathematics	Reading	Word Recognition	Mathematics	Reading	Word Recognition
Grade 1	-0.221 (0.709)	-4.718 (20.207)	-3.101 (6.062)	N/A	N/A	N/A
Grade 2	1.136 (2.875)	0.620 (0.619)	0.329 (0.557)	1.086*** (0.030)	0.818*** (0.021)	1.027*** (0.037)
Grade 3	-0.403 (1.894)	0.833 (1.148)	-0.773 (2.934)	0.962*** (0.025)	0.833*** (0.020)	0.863*** (0.029)

Note: Specifications include school effects, the full history of student demographic (race, gender), free lunch status, class size, and teacher characteristics (race, gender, years of experience and highest education level completed). Standard errors in parentheses are clustered at the classroom level. \*\*\*, \*\*, \* indicate statistical significance at the 1%, 5%, and 10% level respectively. The sample used for the estimates in each cell consists of the same 2239 students whose characteristics are summarized in Table 1.

Table 3: Comparing Alternative Empirical Approaches to Estimate the Effect of Current Class Size on Current Achievement at Various Grade Levels.

Method and Equation -> Estimated Grade level ↓	IV Estimates of Equation (7)	OLS Estimates of Equation (9)	OLS Estimates of Equation (10)	OLS Estimates of Equation (8)	OLS Estimates of Equation (3)	GLS Estimates of Equation (3)	Estimates of Equation (3) assume $\beta_{iT} = \alpha_{iT-1} = 1$	IV Estimates of Equation (7)
Years of Lagged Inputs Included in the Specification	From Kindergarten to Current Grade	Current Grade Only	Current Grade Only	Current Grade Only	From Kindergarten to Current Grade	From Kindergarten to Current Grade	From Kindergarten to Current Grade	Current and previous grade level only
Mathematics								
Grade 1	N/A	-0.324 (0.262)	-0.791*** (0.235)	-1.356*** (0.247)	-1.26*** (0.309)	-0.962*** (0.200)	-0.746*** (0.234)	N/A
Grade 2	-0.365 (0.299)	0.230 (0.267)	-0.031 (0.245)	-0.816*** (0.264)	-0.490 (0.509)	-0.593** (0.263)	-0.419 (0.452)	-0.399 (0.293)
Grade 3	-0.183 (0.318)	0.119 (0.236)	-0.089 (0.204)	-0.510** (0.233)	0.303 (0.497)	0.259 (0.243)	0.079 (0.298)	-0.182 (0.318)
Reading								
Grade 1	N/A	-0.774*** (0.261)	-0.764*** (0.264)	-1.499 *** (0.285)	-1.105*** (0.349)	-0.759*** (0.262)	-0.773*** (0.265)	N/A
Grade 2	-0.369 (0.277)	0.314 (0.244)	-0.105 (0.208)	-0.760*** (0.247)	-0.427 (0.485)	-0.438* (0.253)	-0.411 (0.410)	-0.380 (0.271)
Grade 3	-0.651** (0.277)	-0.102 (0.205)	-0.316* (0.171)	-0.704*** (0.206)	-0.023 (0.431)	-0.155 (0.233)	-0.530* (0.284)	-0.651** (0.277)
Word Recognition								
Grade 1	N/A	-0.765*** (0.281)	-1.008*** (0.282)	-1.539*** (0.301)	-1.442*** (0.391)	-1.172*** (0.297)	-1.003*** (0.310)	N/A
Grade 2	-0.954** (0.397)	0.180 (0.276)	-0.212 (0.260)	-0.886*** (0.308)	-1.011* (0.614)	-0.918*** (0.342)	-0.980* (0.565)	-1.019** (0.388)
Grade 3	-0.656 (0.413)	-0.061 (0.259)	-0.320 (0.209)	-0.637** (0.249)	0.216 (0.542)	-0.042 (0.358)	-0.898** (0.441)	-0.673* (0.404)

Note: Each cell contains the coefficient estimate of class size on a measure of achievement (panels of the table) at a specific grade level (listed by rows). The method and equation estimated are listed in the first row of the table. The second row lists the years of

lagged inputs included in the specification. Specifications also include school effects, student demographics, free lunch status and teacher characteristics. Specifications differ in the years of lagged inputs included as well as whether lagged achievement is included as a sufficient statistic. Corrected standard errors at the classroom level in parentheses. \*\*\*, \*\*, \* indicate statistical significance at the 1%, 5%, and 10% level respectively. The instruments used for column 1 and 8 correspond to IV set 2 in Table 2. The sample used for the estimates in each cell consists of the same 2239 students whose characteristics are summarized in Table 1.

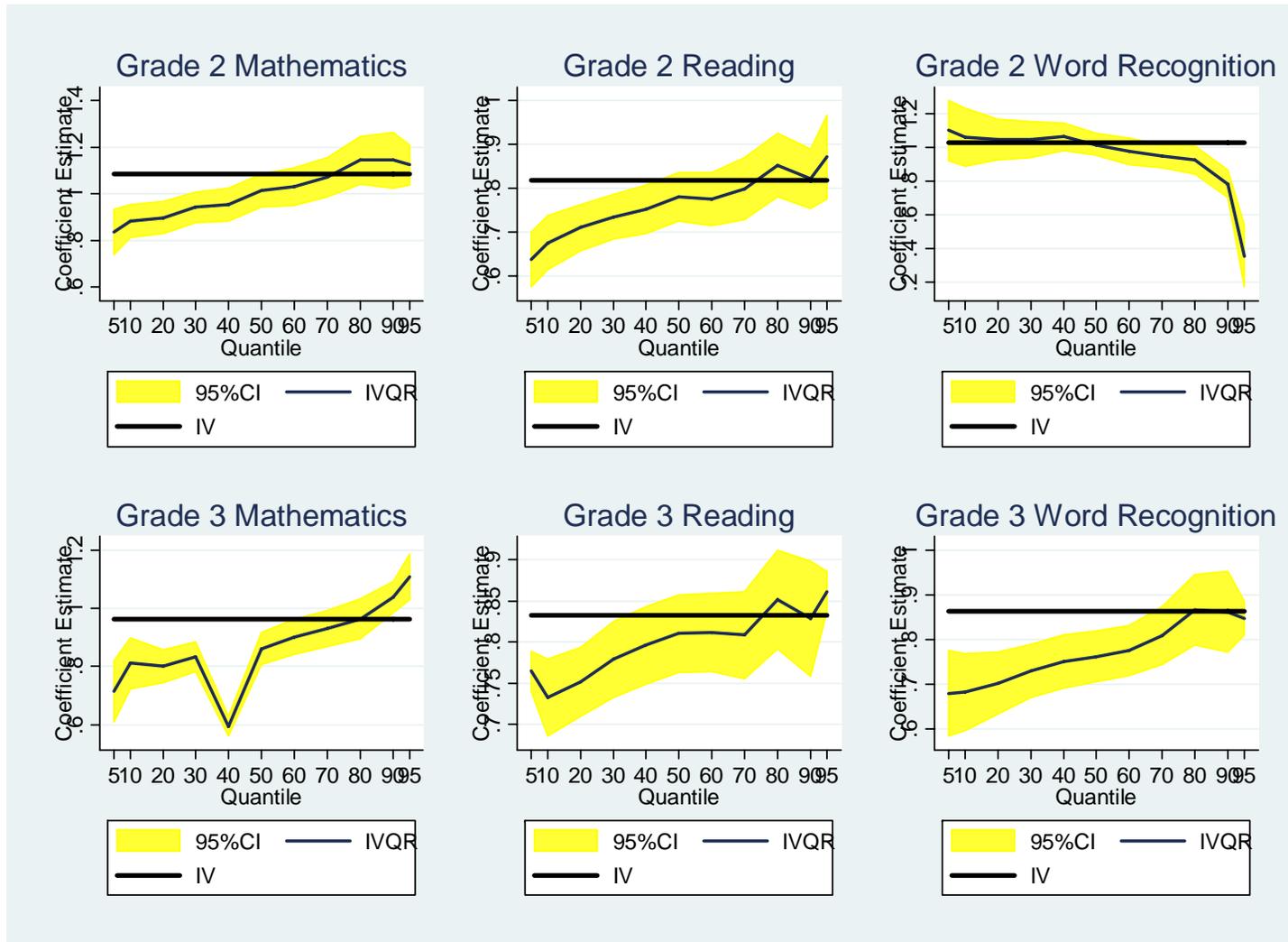
Table 4: Comparing Estimates of Contemporaneous Inputs From IV Estimates of Equations (7)

Years of Lagged Inputs Included in the Specification	From Kindergarten to Current Grade	Current and previous grade level only	From Kindergarten to Current Grade	Current and previous grade level only	From Kindergarten to Current Grade	Current and previous grade level only
Subject	Mathematics		Reading		Word Recognition	
<b>Grade 2</b>						
Unobserved Ability Ratio	1.086*** (0.030)	1.085*** (0.030)	0.818*** (0.021)	0.814*** (0.021)	1.027*** (0.037)	1.022*** (0.037)
Current Class Size	-0.365 (0.299)	-0.399 (0.293)	-0.369 (0.278)	-0.380 (0.271)	-0.954** (0.397)	-1.019** (0.388)
Female Student	2.726** (1.193)	2.668** (1.191)	1.780 (1.118)	1.749 (1.117)	-1.870 (1.589)	-1.878 (1.584)
Student is White/Asian	-5.779** (2.802)	-5.474 (2.793)	1.287 (2.571)	1.507 (2.561)	0.966 (3.704)	1.333 (3.680)
Current Free Lunch Status	-2.340 (2.168)	-3.011 (2.073)	0.196 (2.017)	-0.959 (1.933)	-3.196 (2.884)	-3.853 (2.749)
Current Teacher is Not White	-3.579* (1.990)	-3.576 (1.977)	-3.153* (1.831)	-2.720 (1.818)	-2.848 (2.618)	-2.809 (2.594)
Current Teacher has a Master's Degree	1.449 (1.462)	1.578 (1.460)	-0.624 (1.351)	-0.540 (1.348)	-0.375 (1.943)	-0.162 (1.935)
Teacher Years of Experience	0.039 (0.073)	0.037 (0.073)	0.267*** (0.067)	0.262*** (0.067)	0.326*** (0.093)	0.317*** (0.092)
<b>Grade 3</b>						
Unobserved Ability Ratio	0.960*** (0.025)	0.962*** (0.025)	0.834*** (0.021)	0.833*** (0.020)	0.863*** (0.029)	0.860*** (0.029)
Current Class Size	-0.165 (0.324)	-0.183 (0.318)	-0.602** (0.284)	-0.651** (0.277)	-0.625 (0.423)	-0.656 (0.413)
Female Student	1.570 (1.257)	1.696 (1.260)	1.466 (1.112)	1.503 (1.109)	5.547*** (1.639)	5.371*** (1.634)
Student is White/Asian	-4.980* (3.058)	-4.959* (3.050)	2.125 (2.674)	1.895 (2.653)	6.866* (3.962)	6.600* (3.928)
Current Free Lunch Status	0.837 (2.482)	0.612 (2.374)	-0.766 (2.178)	0.297 (2.071)	2.033 (3.236)	3.376 (3.080)
Current Teacher is	-4.669* (1.990)	-3.595 (1.977)	-0.850 (1.831)	-0.922 (1.818)	-3.001 (2.618)	-3.298 (2.594)

Not White	(2.458)	(2.437)	(2.148)	(2.121)	(3.200)	(3.158)
Current Teacher has a Master's Degree	3.499** (1.579)	3.583** (1.581)	0.802 (1.383)	0.833 (1.378)	-0.019 (2.050)	0.164 (2.043)
Teacher Years of Experience	0.184* (0.086)	0.137 (0.085)	0.178** (0.075)	0.176** (0.074)	0.120 (0.112)	0.137 (0.110)

Note: Corrected standard errors at the classroom level in parentheses. The inputs contained in the specification at each grade level is identical to that in Table 2 and the columns differs in the number of periods of lagged inputs that is listed in the first row. \*\*\*, \*\*, \* indicate statistical significance at the 1%, 5%, and 10% level respectively. The sample used for the estimates consists of the same 2239 students whose characteristics are summarized in Table 1. The instruments for each column correspond to IV set 2 in Table 2.

Figure 1: Instrument Variable and Quantile Regression Instrumental Variable Estimates of the Impacts of Unobserved Ability



Note: Specifications include the full history of inputs listed in Table 2. Two or more lagged test scores are used as instruments.

Appendix Table 1: Impact of the Instruments in the First Stage Regressions

Endogenous Regressor	Grade 1 Mathematics	Grade 1 Reading	Grade 1 Word Recognition	Grade 2 Mathematics	Grade 2 Reading	Grade 2 Word Recognition
<b>IV Set 1 Random Assignment</b>						
Randomly Assigned to Small Class Treatment	-1.479 (4.553)	-5.898 (5.961)	-9.375 (6.273)	-3.718 (5.579)	-3.114 (5.581)	3.608 (6.879)
First Stage F statistic	0.11 [0.745]	0.98 [0.323]	2.23 [0.135]	0.44 [0.550]	0.31 [0.577]	0.28 [0.600]
<b>IV Set 2 Lagged Test Scores</b>						
Kindergarten Mathematics	0.394 (0.017)**	0.199 (0.022)**	0.181 (0.024)**	0.174 (0.021)**	0.063 (0.020)**	0.055 (0.024)*
Kindergarten Reading	0.140 (0.046)**	0.450 (0.061)**	0.334 (0.066)**	0.034 (0.054)	0.084 (0.051)	0.031 (0.061)
Kindergarten Word Recognition	0.090 (0.036)*	0.267 (0.048)**	0.224 (0.052)**	0.001 (0.043)	0.054 (0.040)	0.078 (0.048)
Grade 1 Mathematics	<i>Not included in specification</i>	<i>Not included in specification</i>	<i>Not included in specification</i>	0.464 (0.024)**	0.174 (0.023)**	0.119 (0.027)**
Grade 1 Reading	<i>Not included in specification</i>	<i>Not included in specification</i>	<i>Not included in specification</i>	0.162 (0.030)**	0.374 (0.028)**	0.421 (0.034)**
Grade 1 Word Recognition	<i>Not included in specification</i>	<i>Not included in specification</i>	<i>Not included in specification</i>	-0.015 (0.029)	0.067 (0.028)*	0.121 (0.033)**
First Stage F statistic	464.25 [0.000]	456.33 [0.000]	257.95 [0.000]	281.38 [0.000]	347.81 [0.000]	286.25 [0.000]

Note: Specifications include school effects, current and the full history of student demographic (race, gender), free lunch status, class size, and teacher characteristics (race, gender, years of experience and highest education level completed). Standard errors in () parentheses, Prob >F in [] parentheses. \*\*\*, \*\*, \* indicate statistical significance at the 1%, 5%, and 10% level respectively. Note the grade 1 and grade 2 endogenous regressors listed in the first row are used to identify the time varying impact of unobserved heterogeneity in the specific subject areas respectively in grades 2 and 3 in table 2.