Understanding the Role of Time-Varying Unobserved Ability Heterogeneity in Education Production

Weili Ding
Queen's University

Steven F. Lehrer*
Queen's University and NBER

August 2012

Abstract

Unobserved ability heterogeneity has long been postulated to play a key role in human capital development. Traditional strategies to estimate education production functions do not allow for varying role or development of unobserved ability as a child ages. Such restrictions are highly inconsistent with a growing body of scientific evidence; moreover, in order to obtain unbiased parameter estimates of observed educational inputs, researchers must properly account for unobserved skills that may be correlated with other inputs to the production process. To illustrate our empirical strategy we use experimental data from Tennessee's Student/Teacher Achievement Ratio experiment, known as Project STAR. We find that unobserved ability is endogenously developed over time and its impact on cognitive achievement varies significantly between grades in all subject areas. Moreover, we present evidence that accounting for time-varying unobserved ability across individuals and a more general depreciating pattern of observed inputs are both important when estimating education production functions.

JEL codes: I20, J24, C33 and C81.

*We would like to thank Richard Murnane and seminar participants at Queen's University, Simon Fraser University, 2008 CSWEP/CEMENT workshop, 2008 CEA Annual meetings and the AEA annual meetings session on Education for the Disadvantaged for helpful comments and suggestions. We are grateful to Alan Krueger for generously providing a subset of the data used in the study. 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), hundreds 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 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.

Many classic studies in the economics literature, including Ben-Porath (1967) and Griliches (1977), emphasize that unobserved ability is an input into the production of human capital, but are ambiguous about how they influence human capital accumulation. As a result, within the economics of education literature researchers often use imperfect proxies for unobserved ability or assume their impacts are constant over time or between siblings including twins. These strategies allow 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.[1] These skills are not fixed following conception but rather are related to development of specific brain structures that emerge from both epigenetic and genetic processes. Since unobserved ability heterogeneity is potentially an important contributor to the development of human capital,[2] it would be advantageous to account

[1] 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)).

[2] Within the labor economics literature the empirical importance of unobserved ability heterogeneity to lifetime

for its impacts when estimating education production functions in a more flexible manner than existing methods. More generally, to obtain unbiased parameter estimates of educational inputs researchers must properly control for unobserved ability when estimating education production functions.

Since human capital accumulation is a dynamic processes, it is important to understand how the role of heterogeneous ability evolves over the lifecycle, particularly during periods in which it is most adaptable to policy intervention. To estimate the changing importance of heterogeneous ability differences on academic performance, we introduce a straightforward empirical approach that permits estimation of the time-varying effect of unobserved ability heterogeneity within the standard framework of education production functions.[3] Our empirical strategy exploits the triangular structure implied by the underlying model of human capital production and it is important to state explicitly that this empirical approach does not require measures that either proxy for unobserved ability or make assumptions regarding the process by which unobserved ability develops over the lifecycle.[4] Most importantly, the estimates provide guidance on not only the changing impacts of unobserved ability heterogeneity at both different ages and in different subject areas, but also shed light on how researchers should treat this factor in their analyses.

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 return to educations between 1978 and 1986 for young workers is attributable to a rise in the return to ability. Heckman and Vytlacil (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.

[3]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 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)).

[4]The empirical strategy allows the observed education inputs to both have impacts that vary at different ages and where these inputs could be potentially correlated with the time varying unobserved ability heterogeneity. We discuss the conditions to achieve consistent estimates with both exogenous and endogenous inputs.

To improve our understanding of the importance of unobserved ability heterogeneity in the production of achievement at different ages 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 within schools to classrooms in each year of the experiment to overcome important sources of bias in estimating education production functions, including student-teacher sorting bias (Rothstein (2010)). We empirically demonstrate that it is important to account for the time-varying effects of unobserved individual ability heterogeneity, particularly in reading, listening skills and word recognition. Further, specification tests suggest that this factor should be treated as endogenous in the empirical analysis. While our empirical application is within the economics of education, this empirical strategy could be used in other contexts where unobserved unit-specific heterogeneity is believed to play an important role and may have time-varying impacts. For example, this strategy could be used to estimate whether this source of unobserved heterogeneity accounts for much of the gaps that develop among individuals, groups, countries on outcomes such as health and wealth accumulation.

Similar to Andrabi et al (2011) dynamic panel methods are used to estimate education production functions. However, our empirical strategy differs by exploiting the triangular structure of the underlying economic model of human capital accumulation allowing us to i) provide a structural interpretation of what is often termed the persistence parameter,[5] ii) relax some of the assumptions implicitly made when using a traditional value added estimator, and iii) easily employ semiparametric estimators to explore the extent of student heterogeneity in their endogenous learning rates.[6] We present evidence of substantial heterogeneity in learning rates across students, particularly in mathematics.

[5] Andrabi et al (2011) conclude their investigation by stating that the economic interpretation of the persistence parameter remains an area open for enquiry. This paper is able to provide a clear economic interpretation by exploiting the triangular structure of the underlying economic model.

[6] As we discuss in further detail in the next section, feasible approaches to estimate conditional quantiles with panel data and endogenous regressors are difficult to develop since standard demeaning (or differencing) techniques do not generally remove the time-invariant unobserved heterogeneity. Our approach involves first solving for the unobserved heterogeneity so that estimators based on the L_1-norm penalty can be utilized.

This paper is organized as follows. In Section 2, we review the general conceptual model of cognitive achievement and introduce an empirical strategy that allows for very general patterns of the impacts of both observed and unobserved inputs to the education production process. The estimator requires that a researcher has at least two years of data on education outputs and inputs and identifies the time-varying impacts of unobserved ability heterogeneity using a GMM procedure. We detail the conditions under which this empirical strategy can obtain consistent estimates of the production function parameters, both with and in the absence of ideal data.[7] Project STAR experimental data is described in Section 3. The empirical results that shed light on how researchers should treat unobserved ability heterogeneity are presented and discussed in Section 4. In this section, we also demonstrate that the sign, magnitude and statistical significance of the impact of educational inputs on measures of academic performance is sensitive to restrictions imposed on both unobserved ability heterogeneity and the empirical specification of the education production function. A concluding section summarizes our findings on how researchers should empirically treat unobserved ability heterogeneity in their analyses and discusses direction for future research.

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

[7] As such, the empirical strategy we discuss nests several popular approaches to estimate education production functions. These approaches place implicit and constraining 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) use NLSY79-CS data to investigate the assumptions underlying commonly used achievement production functions (assuming the impact of unobserved ability heterogeneity to be time invariant) and found little empirical support for these assumptions. Our results also complement Andrabi et al (2011) who demonstrate that failing to properly specify and estimate education production functions can yield wildly different results, particularly when there are large gaps in baseline achievement. In Section 4, we conduct model specification tests to determine which (if any) of these alternative assumptions is supported.

function of the full history of family, community, school inputs and own innate ability. These variables may interact with each other in a nontrivial, unknown way. This general model expresses current achievement over time as

$$A_{iT} = f_T(F_{iT}...F_{i0}, S_{iT}...S_{i0}, I_{iT,}\epsilon_{iT}...\epsilon_{i0}), \tag{1}$$

where F_{iT} is a vector of individual and family characteristics, S_{iT} is a vector of school and community characteristics, I_{iT} is a vector of individual current unobserved heterogeneity, including such factors as student innate abilities and determination and ϵ_{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 to assess how specific inputs influence the development of A_{iT}.

2.1 Empirical Cumulative Model

Linearizing the achievement relationship (equation (1)) yields

$$A_{iT} = \beta_{0T} + \beta_{1T}F_{iT} + \beta_{2T}S_{iT} + \beta_{IT}I_i + \sum_{t=0}^{T-1}(\beta_{1t}^T F_{it} + \beta_{2t}^T S_{it} + \rho_t^T \epsilon_{it}) + \epsilon_{iT}. \tag{2}$$

We are essentially extending the traditional panel data model by imposing a multi-factor error structure, where β_{IT} is a vector of factor loadings and I_i corresponds to common unobserved factors.[8] 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 I_i and the idiosyncratic error terms onto the theory. For ease of exposition, we will assume that I_i is a individual scalar. Note, the classical individual effects model that is used in the

[8]For ease of exposition, we will ignore factor dynamics and assume that I_i is a individual scalar fixed over time. Strategies to estimate panel data models with multi-factor error structures are developed in Bai (2009) when the observed covariates are exogenous, and in Harding and LaMarche (2009) for endogenous observed covariates. In this paper, we exploit the triangular structure of the empirical cumulative model of human capital development to estimate how a scalar $\beta_{IT}I_{iT}$ varies in early childhood across subject areas. Our approach can accommodate both exogenous and endogenous observed covariates.

economics of education literature can be obtained by setting $\beta_{IT} = 1$.[9] We place no restrictions on how β_{It} $(t = 0, .., T)$ evolves over time, allowing us to re-express the relationship as

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

where X_{it} is a matrix containing the intercept and all the inputs, $([1, F_{it}, S_{it}]) \, \forall \, t$, that we will assume are independent from ϵ_{iT}.[10] Note that β_t^T represents the matrix of the estimated coefficients that capture how all the inputs from period t affect achievement in period T. Similarly, the relationship in the previous period can be expressed as

$$A_{iT-1} = \beta_{T-1} X_{iT-1} + \beta_{IT-1} I_i + \sum_{t=0}^{T-2} (\beta_t^{T-1} X_{it} + \rho_t^{T-1} \epsilon_{it}) + \epsilon_{iT-1}. \tag{4}$$

Notice the difference in coefficient vectors between equations (3) and (4) as distinguished by superscript T and T-1. We do not impose any restrictions on how the effects of the full set of education inputs on achievement levels varies over time. Further note that the system of equations generated by equations (3) and (4) is triangular in structure. Reexpressing the relationship in equation (4) as a function of unobserved heterogeneity yields:

$$I_i = \frac{1}{\beta_{IT-1}} (A_{iT-1} - \epsilon_{iT-1} - \sum_{t=0}^{T-1} \beta_t^{T-1} X_{it} - \sum_{t=0}^{T-2} \rho_t^{T-1} \epsilon_{it}) \tag{5}$$

Substituting equation (5) into equation (3) yields

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

[9] Variants of this model assuming this classical individual fixed effects structure are also the starting point for analyses of coefficient biases from estimates of education production function. Boardman and Murnane (1979) begin with equation (2), assuming only the current serially uncorrelated residual is included. Todd and Wolpin (2003) include current random shocks but not the full history in equation (2) and assume that shocks are serially correlated. Hanushek (1979) does not include residuals in equation (2).

[10] Later in this section, we discuss how one could estimate education production functions with both time varying ability heterogeneity and endogenous inputs.

where $v_{iT} = \epsilon_{iT} + \sum_{t=0}^{T-1}(\rho_t^T - \frac{\beta_{IT}}{\beta_{IT-1}}\rho_t^{T-1})\epsilon_{it}$ with $\rho_{T-1}^{T-1} = 1$.

Direct OLS estimation of equation (6) will not yield unbiased estimates since A_{iT-1} is correlated with the error term v_{iT}, which contains ϵ_{iT-1}— a component of A_{iT-1}. An instrumental variables (IV) approach can be used to overcome this endogeneity problem and provide consistent estimates of the parameter $\frac{\beta_{IT}}{\beta_{IT-1}}$, 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 $T-2$ or earlier, or differences in lagged test scores.[11] 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. Finally, note if X_{it} contains endogenous inputs to the production process, consistent estimates of these factors could be obtained using the same IV strategy, provided one has access to enough additional instruments to identify both the impacts of these inputs and unobserved ability.[12]

Hausman tests comparing IV and OLS estimates of equation (6) can be used to test for the endogeneity of educational inputs and unobserved heterogeneity. Researchers could also use estimates of equation (6) to conduct specification tests on $\frac{\beta_{IT}}{\beta_{IT-1}}$. Tests on 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.[13]

[11] 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. The moment conditions (using test scores in levels and ignoring other covariates) in this case are given by $E[(A_{iT} - \frac{\beta_{IT}}{\beta_{IT-1}}A_{iT-1})A_{it-j}] = 0 \ \forall j = 2,..,T-1$ and $t = 3,....T$. More generally, if the optimal vector of instruments for period T is denoted by Z_{iT}, then $E[Z_{iT}'v_{iT}] = 0$.

[12] In this situation, the moment conditions are slightly more stringent. As in the previous footnote, if the optimal vector of instruments for period T is denoted by Z_{iT}, then $E[Z_{iT}'\omega_{iT}] = 0$, where we define ω_{iT} to be the matrix consisting of v_{iT} and all the residuals from all of the the first stage equations.

[13] Appendix 1 reviews the three most popular empirical approaches to estimate education production functions, the contemporaneous model, linear growth model and value added model. These approaches are often taken due to data limitations but also ease of implementation. Each approach either implicitly assumes that the impacts of

Even without data on the full history of inputs one can still account for the time-varying impacts of unobserved ability heterogeneity. For example, with only recent data on inputs beginning with period m $(t = m, .., T)$, we can use the same logic that generated equation (6) and express A_{iT} as

$$A_{iT} = \beta_T X_{iT} + \frac{\beta_{IT}}{\beta_{IT-1}} A_{iT-1} + \sum_{t=T-m}^{T-1} (\beta_t^T - \frac{\beta_{IT}}{\beta_{IT-1}} \beta_t^{T-1}) X_{it} + v_{iT}^m \tag{7}$$

where $v_{iT}^m = \varepsilon_{iT} + \sum_{t=0}^{T-m} (\beta_t^T - \frac{\beta_{IT}}{\beta_{IT-1}} \beta_t^{T-1}) X_{it} + \sum_{t=0}^{T-1} (\rho_t - \frac{\beta_{IT}}{\beta_{IT-1}} \rho_t^{T-1}) \epsilon_{it}$ with $\rho_{T-1}^{T-1} = 1$. We can re-express v_{iT}^m in terms of v_{iT} as $v_{iT}^m = v_{iT} + \sum_{t=0}^{T-m} (\beta_t^T - \frac{\beta_{IT}}{\beta_{IT-1}} \beta_t^{T-1}) X_{it}$.

Estimation of equation (7) could also be undertaken via instrumental variables estimation. 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 for which we have no data on. Lagged dependent variables are valid instruments, provided that for some period l $s.t. (T - m < l < T - 1), s.t.$ $\beta_t^T = \frac{\beta_{IT}}{\beta_{IT-1}} \beta_t^{T-1}$ 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 time-varying impacts of unobserved ability heterogeneity, and 2) is somewhat less restrictive in assuming that some past achievement A_{il} is a sufficient statistic for l periods of lagged inputs as opposed to assuming the immediate past achievement A_{iT-1} as a sufficient statistic for all $T - 1$ periods of lagged inputs. Intuitively, this implies that the only way some past inputs could affect current achievement in equation (7) is through the lagged dependent variable.[14] Of course if one has access to exogenous variables other than lagged dependent variables, this assumption about how past inputs enter into the current education production process is unnecessary. As we will discuss in detail in section 4, we consider using information from both random assignment of kindergarten class type from the experiment itself and lagged test scores in other subject areas as instrumental variables. Lastly, it is important to note that it is possible to adopt the Arellano and Bond (1991) specification tests that detect serial correlation in the error term in a dynamic panel data model, where the disturbances are uncorrelated under the Null and follow a moving average process under the alternative. Results from

unobserved ability heterogeneity is fixed as a child ages or that the impact does not exist.

[14] Similar to footnote 10, with multiple endogenous inputs and A_{iT-1}, each element in the set of instruments Z_{iT} is required to be uncorrelated with all of the structural errors in the system of equations not just v_{iT}^m.

these specification tests provide stronger evidence regarding instrument validity than traditional overidentification tests that are known to have poor statistical power.

In our analysis, we consider both OLS and IV estimation of equation (7) and similar to Dewey et al. (2000) we will conduct specification tests to determine if the unobserved ability input to the production process could be treated as exogenous. Further, we will test whether $\frac{\beta_{IT}}{\beta_{IT-1}} = 1$ and consider the consequences of imposing this restriction on both the coefficients and statistical inference of the remaining inputs; we will investigate situations where the data supports this restriction as well as when it refutes it. Finally, we will examine the consequences of imposing $\beta_{it} = 0 \; \forall \; t$, the case where unobserved ability heterogeneity is ignored.

Last, a key advantage of estimating equation (7) in place of alternative equations that account for student unobserved heterogeneity with panel data is that many semiparametric estimators can be used. After all, as with non-linear panel data models, standard demeaning (or differencing) techniques do not result in feasible approaches with conditional quantiles.[15] By substituting equation (5) into equation (3) reduces the dimensions of unobserved variables permitting us to use estimators based on the L_1-norm penalty. In our context, with an endogenous regressor we can use the quantile regression instrumental variables (QRIV) estimator introduced in Cherzonukov and Hansen (2005) to determine if there are potentially different impacts from unobserved ability in different parts of the conditional achievement distribution within grades. Intuitively, using this estimator to recover $\frac{\beta_{IT}}{\beta_{IT-1}}$ provides use with an opportunity to examine whether is heterogeneity in the learning rates over the student population.

3 Data

We use data from Tennessee's highly influential class size experiment, Project STAR to conduct this analysis. 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 kindergarten

[15] See Canay (2011) and the references within for more details on the challenges of estimating quantile regression models with exogenous covariates and panel data.

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). Students who would newly enter the Project STAR schools in later grades were conditionally randomly assigned to class type. In each year of the experiment teachers were also randomly assigned to classrooms.

This dataset has four features which make it ideal to use the empirical strategy described in the preceding section to improve our understanding of the impacts of unobserved ability. First, strictly speaking, one would need data from at least 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.[16] Second, random assignment overcomes selection bias that arises not solely by decisions made by parents, 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.[17] Third, this data set reduces measurement error from aggregation bias by precisely matching each student to the classroom and the teacher within a school, so that we can focus on estimates of the time-varying impacts of individual unobserved ability. Finally, Project STAR was conducted for a single cohort of children between Kindergarten to grade 3, stages in the lifecycle child development specialists have suggested 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 their performance in different dimensions. In this paper, our outcome measures (A_{iT}) are total scaled scores from the Reading, Listening Skills, Mathematics, Word Recognition sections of the Stanford

[16] The standard error or the precision of the estimates may be affected in this case.

[17] Rothstein (2010) 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.

Achievement test.[18] Scaled scores are calculated from the actual number of correct items, 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; within the same test they have the advantage that a 1 point change on one part of the scale is equivalent to a 1 point change on another part of the scale. This scaling offers an important advantage in the identification of $\frac{\beta_{IT}}{\beta_{IT-1}}$, the ratio of the effects of unobserved heterogeneity in between two periods. 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 $\frac{\beta_{IT}}{\beta_{IT-1}}$ 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 in using Project STAR data is that violations to the experimental protocol were prevalent. By grade 3 over 50% of the subjects who participated in kindergarten left the STAR sample and approximately 10% of the remaining subjects switched class type annually. Ding and Lehrer (2010) 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.[19] In order to minimize issues related to the changing composition of the sample that may affect the estimates of unobserved ability heterogeneity, we only include students who participated in all four years of Project STAR and completed exams in all four subject areas each year in this study. Last, an important limitation of Project STAR data is that it contains no information on family inputs beyond free lunch status.

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

[18] 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.

[19] Among this group of students those on free lunch were significantly more likely to be assigned to regular (larger) classes. It should be noted that in 1986 attendance of kindergarten was not mandatory in Tennessee. Thus, students who entered school in grade one may differ in unobservables to those who started in kindergarten.

period. While there are few differences in the percentage of the sample on free lunch across the grades, between grade levels approximately 15% of the students on free-lunch are new recipients. 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. In contrast, 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.[20]

4 Results

In this section, we present evidence that accounting for the time-varying impact of unobserved ability heterogeneity is important. Table 2 shows IV estimates of $\frac{\beta_{IT}}{\beta_{IT-1}}$, the ratio of the effects of unobserved individual heterogeneity from equation (7) using inputs from kindergarten onwards with two alternative instrument sets. As noted earlier, 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.[21] Second, we use two or more periods lagged achievement

[20] These variables are identical to 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 two 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. Note that 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 in the previous academic year.

[21] 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.

scores from all of the other subject areas in the earlier grades.[22] That is, if we are instrumenting for second grade mathematics in equation (7), we can potentially use kindergarten and first grade test scores in the three remaining subject areas. Andrabi et al. (2011) present evidence that by employing test scores in other subject areas, biases from measurement error are reduced relative to using sufficiently lagged test scores in the same subject. To formally examine the validity of the exclusion restriction with these lagged test scores as instruments, we will conduct a simple modification of the Arellano and Bond (1991) m_2 specification tests to see if lagged residuals in other subjects areas are sufficiently correlated with residuals in equation (7).[23]

Employing only the initial random assignment to class type as an instrument provides imprecise and statistically insignificant estimates of $\frac{\beta_{IT}}{\beta_{IT-1}}$ in each subject and grade. The sign and magnitude of $\frac{\beta_{IT}}{\beta_{IT-1}}$ varies substantially with this instrument. Examination of the first stage regression presented in the left panel of Table 3, demonstrates that random assignment to a small class is a very weak instrument for every grade-subject area measure of A_{iT-1}.

IV estimates of equation (7) with the same education inputs but using two or more periods of lagged achievement scores in other subject areas as instruments, provide more precise evidence on $\frac{\beta_{IT}}{\beta_{IT-1}}$. As indicated in the right panel of table 2, the estimated time-varying impact of unobserved ability heterogeneity appears fairly constant across grades in both mathematics and word recognition. In mathematics, the contribution of unobserved ability declines slightly (approximately 11%) between grades 2 and 3. However, in grade 3, the constraint that $\frac{\beta_{IT}}{\beta_{IT-1}} = 1$ is supported. Results from t-tests of this constraint appear in the lower panel of Table 2. With IV set 2, this constraint is firmly rejected in both grades 2 and 3 reading, grade 2 mathematics and both grade 3 word recognition and listening skills tests. Whereas the estimated magnitude of the time-varying impact in reading changes little across grade levels, the estimates suggest (on average) a declining role (< 1) for this factor. Overall, the positive impact of unobserved heterogeneity declines by approximately

[22] The second instrument set could expand in higher grades as more past test scores become available to serve as additional instruments, which presents 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.

[23] Results from these specification tests are presented and discussed later in this subsection.

10.66% between grades one to three in mathematics. In word recognition and listening skills the estimated magnitude of the time-varying effect respectively declines by 17.1% and 9.2% between grades 2 and 3.[24] On average, the estimated impact of unobserved ability heterogeneity on test scores in all subject areas declines (on average) in grade 3 from grade 2.

IV estimates that use lagged test scores as instruments strongly reject the assumption that in grades 2 and 3 unobserved heterogeneity has no effect (i.e. $\beta_{IT} = 0$) in all subject areas. It is important to note that the popular methods used to estimate education production function (reviewed in Appendix 1) assume that (if non-zero) the contemporaneous effects of unobserved heterogeneity are fixed. The results presented in table 2 indicate that this implicit assumption would only be satisfied in the subject area of mathematics with Project STAR data.

Intuitively, these empirical results confirm ideas from the education literature that students require more cumulative knowledge (i.e. literacy) in reading, listening and word recognition, and therefore less reliant on unobserved "ability". While acquiring mathematics knowledge is also a gradual process, the structure of test questions changes sharply from one grade to another. Mathematics tests in grades 2 and 3 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 in different subjects.

In order to examine the importance of accounting for unobserved ability heterogeneity we next calculated the partial R-squared for this variable. The partial R-squared ranged between 20 and 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 compared to what is explained by the full set of current and past observed education inputs.

Table 3 indicates that weak instruments are not a concern for our instrument set containing lagged test scores from other subject areas. The first stage F-statistic of the hypothesis that the coefficients on the excluded instruments are all 0, range from 188.35 to 593.16, with a p-value

[24] Note the results are robust to using a single two-period lagged test score in the same subject area as an instrument.

less than 0.01 in all cases. Additionally, the individual coefficient on most of the instruments is significant at the 1% level, indicating that there is a strong first stage relationship with IV set 2.

To further assess the validity of IV set 2 (and the corresponding moment restrictions), we conducted tests of second-order serial correlation in the residuals from the full system of equations. The test statistic, which is distributed $N(0,1)$, and the p-value of the hypothesis that residuals are serially uncorrelated are reported by instrument endogenous regressor pair in Appendix Table 1. There is little evidence of higher order serial correlation in the residuals. There were only four exceptions where there is some failure at the 10% level and these instruments were removed from the specification.[25] Given few differences in the raw correlation between the outcome variables where failures of this test did or did not occur, we could not devise a rule of thumb that might predict when these failures are most likely to happen. Overall, the tests in Appendix Table 1 generally reject that there is second (or third) order serial correlation in the residuals, increasing our confidence that the statistical properties of the instruments are met.

4.1 How should unobserved ability be treated?

To provide further guidance on how researchers should treat unobserved ability heterogeneity, we consider several alternative strategies to estimate equation (7). Table 4 illustrates how the sign, magnitude and statistical significance of the estimated coefficient on four contemporaneous inputs (class size, student race, gender and free lunch status) from equation (7) differ based on how one treats $\beta_{IT} I_{iT}$ in equation (3). In the first two columns of Table 4, we present IV and OLS estimates of equation (7), where we do not imposes restrictions on the effects of unobserved inputs to the production process and respectively treat A_{iT-1} as endogenous and exogenous. The first column contains IV variable estimates where the lagged achievement scores are used to identify $\frac{\beta_{IT}}{\beta_{IT-1}}$ and this is our preferred specification. In column 3, we impose the restriction that $\frac{\beta_{IT}}{\beta_{IT-1}} = 1$, that unobserved ability heterogeneity has the same effect in all periods when estimating equation (7). In column 4, we consider the consequences from omitting unobserved ability heterogeneity, that is,

[25] Since all of the specifications we considered are over-identified, we did replicate the full analysis where we did not drop these invalid instruments and found that the full set of results is robust to their inclusion.

$\beta_{It} = 0 \; \forall \; t$. In column 5, we consider the consequences of using fewer lagged years of observed inputs in the specification of equation (7) relative to column 1. We employ the same instruments and estimator in columns 1 and 5 of Table 4, but as noted in Section 2, since we include fewer years of lagged inputs as controls, the exclusion restriction assumption may require greater defense in column 5. We examine these specifications for each subject area and grade level.

Estimates of equation (7) between columns 1 and 2 differ in whether A_{iT-1} is treated as endogenous. Not surprisingly, given the evidence in Nickell (1981), OLS estimates of $\frac{\beta_{IT}}{\beta_{IT-1}}$ are downward biased in all grades and subject areas when school fixed effects are included in the specification. Further, Hausman tests between OLS and IV estimates of equation (7) presented in Appendix Table 3 reject both the Null of exogeneity for the entire coefficient vector in all subject areas as well as $\frac{\beta_{IT}}{\beta_{IT-1}}$ by itself in grades 2 and 3. Thus, to estimate education production functions with Project STAR it is necessary to account for individual specific unobserved heterogeneity that is correlated with the full set of inputs.

In column 3 of Table 4, we restrict unobserved heterogeneity to have a constant effect between successive grades, which may be correlated with educational inputs. For math in grade 3, grade 2 listening and word recognition exam, we use estimates from column 1 which suggest that we cannot reject the restriction that $\frac{\beta_{IT}}{\beta_{IT-1}} = 1$ (see lower panel of Table 2 for test results). For these grade-subject pairs, although imposing the restriction leads to little difference in the coefficient estimates, this restriction does increase the residual variation, resulting in larger standard errors. In fact, the standard errors in column 3 are either approximately the same size or larger than the IV estimates in column 1.[26] The constant effect assumption is clearly rejected for reading in grade 2 and 3, for grade 2 math l and 3 word recognition and listening skills. There are several changes in the estimated coefficient on contemporaneous class size and it is generally larger in magnitude than column 1. As before, the standard errors on the other inputs increase in size relative to column 1. Further, when restricting $\frac{\beta_{IT}}{\beta_{IT-1}} = 1$, the impact of class size becomes statistically significant on the grade 3 word recognition exam and is roughly 40% larger in magnitude on the grade 3 reading exam.

[26] This is worth noting since the standard errors for IV estimates are always larger than those obtained by using OLS with the same specification; otherwise the denominator in the Hausman test would be undefined.

On the grade 2 reading exam, the coefficients on the student characteristics differ substantially in magnitude from those presented in columns 1 and 2. Taken together, these results indicate that the constant effect assumption, even when valid, could affect statistical inference of the education input estimates and when clearly rejected could lead to very different results.

Placing restrictions on how researchers treat unobserved ability heterogeneity in their empirical analyses is most troubling in situations where they ignore its role, when the data suggests it significantly affects achievement. The consequences of this restriction are demonstrated in column 4 of Table 4. The coefficients on many student characteristics such as race or free lunch status increase sharply relative to those presented in column 1. This increase suggests that differences in ability heterogeneity account for a large portion of the gap across ethnic and income groups. Estimates on the impact of current class sizes often differ in significant ways between columns 1 and 4 of table 4. For instance, ignoring the role of ability would now suggest that small classes boost achievement on both grade 3 mathematics and grade 2 listening tests, but they are no longer effective for grade 3 reading and word recognition. The large differences in the estimated magnitude of these coefficients between columns 1 and 4 could have large importance for public policy but should not be a surprise since tests that $\frac{\beta_{IT}}{\beta_{IT-1}} = 0$ using estimates from either columns 1 and 2 strongly reject this restriction for any grade and subject level.

Estimates of $\frac{\beta_{IT}}{\beta_{IT-1}}$ as well as contemporaneous home and school inputs on achievement barely exhibited any differences between columns 1 and 5 of Table 4 for each grade-subject pair,[27] only the sign but not statistical significance of the impact of current free lunch status on reading in both grades 2 and 3 changed. This difference arises from the fact that kindergarten free lunch status has a large effect on grade 3 reading, so its exclusion in column 5 leads to omitted variable bias since free lunch status is highly correlated across grades. If one has access to more lagged years of observed inputs, it can reduce concerns related to omitted variable bias and increase the plausibility of using lagged tests scores as instruments. Our results using Project STAR data suggest that in practice the data requirements to estimate equation (7) may not be difficult to satisfy.

[27] Recall that columns 1 and 5 differ solely in the number of lagged inputs in the specification.

4.2 How much unobserved student-level heterogeneity is there in learning?

Figure 1 presents QRIV estimates of $\frac{\beta_{IT}}{\beta_{IT-1}}$ 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 $\frac{\beta_{IT}}{\beta_{IT-1}}$. 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. For instance, in mathematics the impact at higher deciles is 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.[28] The gap between an individual at the highest quantile in both grades relative to an individual at the lowest quantile is over 115%.

Figure 1 suggests that even at early ages in school, large differences in the impacts of unobserved ability heterogeneity across the population appear. 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, listening skills and word recognition. While Table 1 reported that the impact of unobserved ability was on average not significantly different from one in mathematics, Figure 1 presented substantial heterogeneity in these estimated impacts across the distribution. This heterogeneity further demonstrates that if rank invariance holds then traditional differencing approaches of education production functions may not be appropriate since for individuals at many quantiles unobserved ability does not evolve at a constant

[28] With the exception of grade two word recognition individuals at higher deciles generally experience larger impacts from unobserved ability heterogeneity. 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.

rate. 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 societal 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.

4.3 Comparing education production function specifications

Estimating equation (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 restrictions that the method imposes on how the impact of past observed education inputs decay compared to the popular approaches detailed in Appendix 1. Similar to Todd and Wolpin, when we compare models of education production based on the amount of historical inputs included, we do not find evidence to support restrictive models, which assume test scores depend only on contemporaneous inputs. In this subsection, we reinforce the findings from earlier research that lagged observed inputs matter in the production of current achievement and that the impact of different inputs decay at different rates, but it is important to repeat that the data also suggests that a more general treatment of how unobserved heterogeneity affects achievement is required.

Since the impact of unobserved ability heterogeneity varies between grades for all subjects with the exception of mathematics, this implies that researchers should be cautious in pooling data on student achievement across grade levels when estimating education production functions. If unobserved ability heterogeneity has differential impacts at different ages and if this factor is correlated with included inputs then further biases may be introduced by restricting it to have a common impact. Further, as shown in Table 4, the coefficients on contemporaneous inputs vary significantly between grades 2 and 3 indicating that specifications that restrict contemporaneous inputs to have the same impact on contemporaneous achievements at different grade levels are highly restrictive.[29]

In order to see which of the empirical specification is most appropriate for equation (1), we reestimate equation (7) by limited information maximum likelihood (LIML) allowing us to con-

[29] Specification tests strongly reject empirical models that restrict the impacts of contemporaneous inputs to have the same effect on both grades 2 and 3 achievement levels.

duct model specification tests between this model and alternative specifications of the education production function.[30] We examine the less flexible methods common in the economics of education literature that are reviewed in Appendix 1.[31] Appendix Table 2 presents likelihood ratio test statistics and their p-values from tests that compares the alternative nested specifications of the education production function. All of the p-values are well below 0.05, indicating that the restrictions of each approach reviewed in Appendix 1 are soundly rejected. Reinforcing this finding, it is worth pointing out that these differences in the specification of the education production function are highly relevant in practice as placing restrictions on the impacts of observed and unobserved inputs to the education production function leads to substantially different estimates and policy recommendations. In particular, estimates of the contemporaneous model (equation (8)) suggest

[30] LIML places additional distributional assumptions on the residual of equation (7), but it is asymptotically equivalent to the GMM strategy assuming homoskedastic and serially uncorrelated errors. While LIML is much less susceptible to weak instruments problem than 2SLS, it could result in drastically different estimates if the residuals are not Normally distributed. As these distributional assumptions are unattractive, we only consider this method for the purpose of conducting these specific model specification tests. Note that the LIML estimates of $\frac{\beta_{IT}}{\beta_{IT-1}}$ did not differ substantially (i.e. less than an order of 3%) from the GMM estimates presented in Tables 2 and 4, and are available upon request.

[31] 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 of all past observed and unobserved inputs $\beta_t^T = 0 \ \forall t \in [0..T-1]$ and $\beta_{IT} = 0$, ii) the linear growth model uses gains in test scores as a dependent variable and assumes that the effects of all past observed and unobserved inputs do not decay, $\beta_t^T = \beta$ $\forall t$, $\beta_{IT} = \beta_{IT-1}$, 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. For comparison, 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) assuming I_{IT} is exogenous, which presents an opportunity to conduct a variety of simple specification tests. Using Wald tests that compare the more general model with one that is restricted and nested within the first model, we found that in all grades and subject areas, the restrictions that underlie each of the three empirical approaches described in Appendix 1 are rejected. Further, the results from grade 3 suggest that while A_{iT-1} is not a good sufficient statistic, A_{iT-2} indeed shows promise. Last, we also conducted model specification tests using the Akaike information criterion (AIC) methods and reached the same conclusions as those presented in Appendix Table 2.

that in all subject areas there is a large statistically significant benefit from reduced class sizes whereas the estimated impact of current class size on achievement from equation (10)) is opposite in sign to that presented in column 1 for all grade 2 subjects and for math in grade 3.[32]

5 Conclusion

In the economics of education literature researchers often implicitly assume that both the impact and stock of unobserved ability are constant over time when estimating education production functions. This appears inconsistent with a rapidly growing body of scientific evidence which indicates that the impacts and development of these unobserved factors vary substantially over the lifecycle. In this paper, we propose specifications of an education production functions that allow for both time-varying unobserved ability within individuals and a more general decaying pattern of past inputs in an effort to improve our understanding of the role of time-varying unobserved ability heterogeneity in the education production process. We present evidence that accounting for both observed inputs and unobserved heterogeneity in a more flexible manner is both appealing and important empirically. Our results suggest that unobserved ability is correlated with observed inputs to the production process. The impacts of unobserved ability on achievement between grade 1 to grade 3 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 3 grade levels even in the same subject area, traditional differencing approaches of education production functions such as the within individual transformation may be invalid. Further, our results indicate that when estimating education production functions with data from multiple grade levels, researchers should be cautious about pooling data, which places unsupported restrictions on how contemporaneous inputs affect achievement measures. Finally, the impacts of unobserved ability vary substantially over the population particularly in mathematics.

Our analysis further supports earlier research that demonstrates how the different empirical

[32] Estimates from the value added, linear growth and contemporaneous models of education production are available upon request.

approaches that are used to estimate education production functions can present substantially different pictures of the effectiveness of inputs such as smaller classes. Thus, 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.[33] 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 quite restrictive and may further bias parameter estimates of observed educational inputs. The results that the impact of unobserved ability differs across subjects also has important implications for accountability and policies that both reward and make retention decisions for teachers based on value added. After all, if one were to ignore student unobserved ability heterogeneity in the analyses and if this factor has differential time-varying effects across subject areas as with the Project STAR data, then the resulting ordering of teacher effects may not reflect teaching quality but rather captures the nature of the subjects taught.

Although our empirical results may not generalize universally, they suggest that researchers should consider adopting more general estimation strategies that place fewer restrictions on the underlying model of education production, particularly given the increasing number of rich longitudinal education datasets being made available around the world. Since the empirical strategy introduced in the paper exploits the triangular structure of the underlying model to identify the impact of time-varying unobserved ability heterogeneity, it may extend beyond education production functions and have implications for empirical researchers that seek to explain cumulative gaps between groups or countries such as growth or wealth as well as those working with other cumulative models of individual human capital 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

[33] 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 (2010)).

these components is time-varying. Developing estimable education production functions from the underlying economic model that adopts recent econometric methods which assume that the unobservable individual effects has a factor structure (i.e. Bai (2009), Harding and Lamarche (2011) and Ahn et al. (2007)) could potentially lead to new policy relevant insights. For instance it may allow us to identify the time-varying impacts of different dimensions of unobserved abilities (i.e. cognitive vs. non cognitive) as well as observed inputs on measures of academic performance to shed light on which targeted education interventions could yield the largest returns.

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] Andrabi, T., J. Das, A. I. Khwaja, and T. Zajonc (2011) "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics," *American Economic Journal: Applied Economics* 3(3), 29–54

[3] 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.

[4] Bai, J., (2009), "Panel Data Models with Interactive Fixed Effects," *Econometrica* 77(4), 1229-1279.

[5] Ben-Porath, Y. (1967), "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy* 75(4), 352-365.

[6] 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.

[7] Canay, I. A. (2011), "A Simple Approach to Quantile Regression for Panel Data," *The Econometrics Journal* 14 (3), 368-386.

[8] 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.

[9] 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.

[10] Chernozhukov, V., and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245-261.

[11] Chernozhukov, V. (2000) Conditional Extremes and Near-Extremes: Estimation, Inference, and Economic Applications," Ph.D. Dissertation, Stanford University.

[12] 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).

[13] Cunha, F. and J. J. Heckman (2008), "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Journal of Human Resources* 43(4), 738-782.

[14] Cunha, F., J. J. Heckman, and S. M. Schennach (2010), "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica* 78(3), 883–931.

[15] 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

[16] Dewey J., T. A. Husted and L. W. Kenny (2000), "The ineffectiveness of school inputs: a product of misspecification?," *Economics of Education Review* 19(1), 27–45.

[17] Ding, W. and S. F. Lehrer (2010), "Estimating Treatment Effects from Contaminated Multi-Period Education Experiments: The Dynamic Impacts of Class Size Reductions," *Review of Economics and Statistics* 92(1), 31-42.

[18] Figlio, D. N., (1999), "Functional Form and the Estimated Effects of School Resources," *Economics of Education Review* 18(2), 241-252.

[19] Griliches, Z., (1977), "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica* 45(1), 1-22.

[20] 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.

[21] 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.

[22] Hanushek, E. A., (1979), "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," *Journal of Human Resources* 14(3), 351-388.

[23] Harding, M. and C. Lamarche (2011), "Least Squares Estimation of a Panel Data Model with Multifactor Error Structure and Endogenous Covariates," *Economics Letters* 111(1), 197-199.

[24] Heckman, J. J. and E. Vytlacil (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.

[25] Heckman, J. J., (2000), "Policies to Foster Human Capital," *Research in Economics* 54(1), 3-56.

[26] 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.

[27] Keane, M. P. and K. I. Wolpin (1997), "The Career Decisions of Young Men," *Journal of Political Economy* 105(3), 473-522.

[28] Krueger, A. B., (1999) "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics* 114(2), 497-532.

[29] Leibowitz, A., (1974), "Home Investments in Children," *Journal of Political Economy* 82(2), S111–131.

[30] 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.

[31] Nickell, S. (1981), "Biases in Dynamic Models with Fixed Effects," *Econometrica* 49(6), 1417-1426.

[32] Rothstein, J., (2010), "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics* 125(1), 175-214.

[33] Shonkoff, J. and D. Phillips eds. (2000), "*From Neurons to Neighborhoods: The Science of Early Childhood Development*," Washington DC: National Academy Press.

[34] 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.

[35] 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.

[36] 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.

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.[34] 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

$$A_{iT} = \beta' X_{iT} + \varepsilon^c_{iT}, \tag{8}$$

where $\varepsilon^c_{iT} = \beta_{IT} I_i + \sum_{t=0}^{T-1} \beta^T_t X_{it} + \epsilon_{iT}$. Unbiased parameter estimates from equation (8) require that past inputs to the production process and unobserved ability decay immediately.[35]

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:

$$A_{ijT} = \beta_T X_{ijT} + \delta A_{ijT-1} + \varepsilon^L_{ijT} \tag{9}$$

where $\varepsilon^L_{iT} = \varepsilon_{iT} + (\beta_{IT} - \delta\beta_{IT-1}) I_i + \sum_{t=0}^{T-1} (\beta^T_t - \delta\beta^{T-1}_t) X_{it}$. The inclusion of A_{iT-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 both observed and

[34] We provide a brief review below and guide the reader to Todd and Wolpin (2003) for a more comprehensive discussion.

[35] This requires $\beta^T_t = 0 \vee t \in [0..T-1]$ 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.

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}.[36]

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 ($\Delta A_{iT} = A_{iT} - A_{iT-1}$),[37] as

$$\Delta A_{iT} = \beta' X_{iT} + \widetilde{\varepsilon}_{iT} \tag{10}$$

where $\widetilde{\varepsilon}_{iT} = \varepsilon_{iT} + (\beta_{IT} - \alpha_{IT-1})I_i + \sum_{t=0}^{T-1}(\beta_t^T - \beta_t^{T-1})X_{it} + \sum_{t=0}^{T-1}(\rho_t - \varsigma_t)\varepsilon_{it}$. 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.[38]

[36] 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.

[37] 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.

[38] This assumption is fairly restrictive as it implies that having a good second grade math teacher has the same impact on an achievement meausre 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

$$\Delta A_{iT} = \beta' X_{iT} + \gamma_I' I_i + \widetilde{\widetilde{\varepsilon}}_{iT} \tag{11}$$

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).

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

	Kindergarten	Grade One	Grade Two	Grade 3
Class Size	19.9079 (3.8279)	20.3323 (4.0179)	20.2015 (4.1943)	20.3931 (4.4458)
Receiving Small Class Treatment	0.3137 (0.4641)	0.3137 (0.4641)	0.3137 (.4641)	0.3137 (0.4641)
Math Test Score	500.0545 (45.1513)	545.9033 (40.4594)	594.5833 (43.5603)	628.0118 (40.0972)
Reading Test Score	445.7054 (31.506)	541.8566 (52.4381)	599.4453 (43.3322)	625.6645 (37.0832)
Word Recognition Test Score	443.7236 (37.3205)	532.8634 (46.8292)	600.0785 (46.9727)	622.8652 (43.8917)
Listening Test Score	546.3895 (31.607)	577.621 (33.0834)	604.1943 (34.2712)	629.5511 (31.0411)
Free Lunch Status	0.3565 (0.4791)	0.3681 (.04824)	0.3522 (0.4778)	0.3499 (0.477)
Student is White of Asian	0.7553 (0.43)	0.7553 (0.43)	0.7553 (0.43)	0.7553 (0.43)
Student is Female	0.5207 (0.4997)	0.5207 (0.4997)	0.5207 (0.4997)	0.5207 (0.4997)
Teacher Race is Non-White	0.1256 (0.3315)	0.1389 (0.3459)	0.1744 (0.3796)	0.1626 (0.369)
Teacher has a Masters Degree	0.3802 (0.4855)	0.3432 (0.4749)	0.3621 (0.4807)	0.4452 (0.4971)
Teacher Years of Experience	9.4701 (5.5013)	11.6936 (8.6052)	13.0882 (8.5536)	13.5635 (8.4419)
New Teacher	0.0689 (0.2534)	0.0908 (0.2874)	0.074 (0.2619)	0.0548 (0.2276)

Note: Each cell reports the mean and standard deviations in parentheses. There are 2203 students who participated and completed all four 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

	IV SET 1 Random Class Type Assignment				IV SET 2 Two or More Period of Lagged Test Scores in Other Subject Areas			
Subject Area	Mathematics	Reading	Word Recognition	Listening Skills	Mathematics	Reading	Word Recognition	Listening Skills
Grade 1	-0.221 (0.709)	-4.718 (20.207)	-3.101 (6.062)	0.223 (0.418)	N/A	N/A	N/A	N/A
Grade 2	0.998 (2.056)	0.700 (0.658)	0.308 (0.506)	7.020 (39.669)	1.078 (0.033)***	0.809 (0.025)***	1.021 (0.033)***	1.005 (0.036)***
Grade 3	-0.633 (3.640)	2.795 (9.832)	-0.857 (3.141)	0.106 (0.960)	0.963 (0.029)***	0.841 (0.023)***	0.846 (0.026)***	0.906 (0.029)***
Results from two sided Wald tests of the Null that $\beta_{iT}/\beta_{iT-1}=1$								
Grade 1	3.39 (0.066)*	0.07 (0.790)	0.46 (0.497)	3.47 (0.063)*	N/A	N/A	N/A	N/A
Grade 2	0.077 (0.946)	0.60 (0.439)	2.46 (0.117)	0.02 (0.901)	5.61 (0.018)**	56.14 (0.000)***	0.40 (0.528)	0.08 (0.771)
Grade 3	0.11 (0.735)	0.01 (0.907)	0.30 (0.587)	0.83 (0.361)	2.42 (0.120)	54.88 (0.000)***	42.22 (0.000)***	9.89 (0.002)***

Note: In the top panel, 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. In the bottom panel, the chi squared-statistic and in parentheses the associated p-value from a two sided test that the parameter equals one are presented

Table 3: Impact of the Instruments in the First Stage Regressions

Endogenous Regressor	Grade 1 Mathematics Test Score	Grade 1 Reading Test Score	Grade 1 Word Recognition Test Score	Grade 1 Listening Skills Score	Grade 2 Mathematics Test Score	Grade 2 Reading Test Score	Grade 2 Word Recognition Test Score	Grade 2 Listening Skills Score
IV Set 1 Random Assignment								
Randomly Assigned to Small Class Treatment	-1.846 (4.369)	-5.582 (6.084)	-10.477 (6.679)	7.068 (3.193)**	-2.480 (6.190)	-1.399 (6.049)	3.630 (7.181)	-3.834 (5.274)
First Stage F statistic	0.18 [0.676]	0.864 [0.353]	2.78 [0.095]	0.04 [0.025]	0.15 [0.683]	0.09 [0.821]	0.00 [0.610]	0.56 [0.375]
IV Set 2 Lagged Test Scores								
Kindergarten Mathematics Score	*Not included in specification*	*Not included due to Appendix table 2 tests*	0.177 (0.026)***	0.230 (0.018)**	*Not included in specification*	0.038 (0.022)	0.027 (0.022)	0.063 (0.019)**
Kindergarten Reading Score	0.301 (0.053)**	*Not included in specification*	0.583 (0.038)***	0.264 (0.043)**	0.123 (0.066)**	*Not included in specification*	0.073 (0.030)*	0.232 (0.046)**
Kindergarten Word Recognition Score	0.089 (0.040)*	0.660 (0.025)***	*Not included in specification*	-0.019 (0.033)	0.009 (0.052)	0.167 (0.025)**	*Not included in specification*	-0.089 (0.038)*
Kindergarten Listening Skills Score	0.370 (0.022)**	0.186 (0.027)***	*Not included due to Appendix table 2 tests*	*Not included in specification*	0.216 (0.032)***	0.088 (0.028)**	0.102 (0.032)**	*Not included in specification*
Grade 1 Mathematics Score	*Not included in specification*	*Not included in specification*	*Not included in specification*	*Not included in specification*	*Not included in specification*	0.183 (0.027)**	0.082 (0.030)**	0.239 (0.023)**
Grade 1 Reading Score	*Not included in specification*	*Not included in specification*	*Not included in specification*	*Not included in specification*	0.283 (0.034)***	*Not included in specification*	0.521 (0.022)**	0.130 (0.015)**
Grade 1 Word Recognition Score	*Not included in specification*	*Not included in specification*	*Not included in specification*	*Not included in specification*	*Not included due to Appendix table 2 tests*	0.358 (0.019)**	*Not included in specification*	*Not included due to Appendix table 2 tests*
Grade 1 Listening Skills Score	*Not included in specification*	*Not included in specification*	*Not included in specification*	*Not included in specification*	0.216 (0.034)***	0.121 (0.036)**	0.060 (0.035)	*Not included in specification*
First Stage F statistic	327.98 [0.000]	593.16 [0.000]	397.43 [0.000]	239.24 [0.000]	204.04 [0.000]	274.03 [0.000]	330.45 [0.000]	188.35 [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. Appendix table 2 test results refer to tests of serial correlation in the residuals of the education production function.

Table 4: Comparing Estimates of Education Inputs Estimates of Equations (7) under Different Assumptions Regarding Education Production

Method	IV Estimation	OLS Estimation	Constrained OLS Estimation	Constrained OLS Estimation	IV Estimation
Years of Lagged Inputs Included in Specification	From Kindergarten	From Kindergarten	From Kindergarten	From Kindergarten	From Grade One
Grade 2 Mathematics					
Unobserved Ability Ratio	1.078 (0.033)***	0.749 (0.021)	β_{iT}=1 is assumed	β_{iT}=0 is assumed	1.077 (0.033) ***
Current Class Size	-0.388 (0.317)	-0.321 (0.427)	-0.398 (0.468)	-0.518 (0.524)	-0.417 (0.293)
Female Student	2.625 (1.194)**	2.316 (1.111)**	2.560 (1.192)**	1.734 (1.534)	2.569 (1.189)**
Student is White/Asian	-5.546 (2.827)**	0.944 (2.454)	-4.139 (2.704)	13.820 (3.415)***	-5.239 (2.792)*
Current Free Lunch Status	-2.434 (2.087)	-4.168 (1.778)**	-2.974 (1.920)	-9.869 (2.468)***	-3.120 (2.069)
Grade 1 Class Size	0.911 (0.365)**	0.515 (0.488)	0.824 (0.528)	-0.287 (0.614)	0.915 (0.322)***
Kindergarten Class Size	-0.022 (0.272)	-0.066 (0.273)	-0.028 (0.291)	-0.105 (0.344)	Not included in specification
Grade 1 Free Lunch	-1.480 (2.172)	-4.500 (2.027)* *	-2.227 (2.128)	-11.768 (3.161)***	-2.461 (2.100)
Kindergarten Free Lunch	-2.273 (1.996)	-4.067 (1.726)**	-2.515 (1.880)	-5.605 (2.437)**	Not included in specification
Grade 3 Mathematics					
Unobserved Ability Ratio	0.963 (0.029)***	0.662 (0.015)***	β_{iT}=1 is assumed	β_{iT}=0 is assumed	0.969 (0.028) ***
Current Class Size	-0.121 (0.318)	-0.054 (0.296)	-0.149 (0.419)	-0.578 (0.288)**	-0.162 (0.318)
Female Student	1.522 (1.262)	1.400 (1.154)	1.483 (1.315)	1.383 (1.106)	1.598 (1.266)
Student is White/Asian	-5.072 (2.780)**	-1.502 (2.796)	-5.489 (2.883)*	2.217 (2.573)	-5.040 (3.064)*
Current Free Lunch Status	0.841 (2.304)	-0.988 (2.273)	1.380 (2.245)	-0.449 (2.015)	0.973 (2.388)
Grade 2 Class Size	0.381 (0.419)	0.101 (0.365)	0.425 (0.569)	0.769 (0.355)**	0.214 (0.321)
Grade 1 Class Size	-0.501 (0.424)	-0.469 (0.356)	-0.539 (0.610)	-0.194 (0.362)	Not included in specification
Kindergarten Class Size	0.262 (0.277)	0.224 (0.259)	0.310 (0.304)	-0.274 (0.251)	Not included in specification
Grade 2 Free Lunch	0.802 (2.510)	-0.311 (2.399)	0.786 (2.470)	-1.904 (2.197)	-0.559 (2.362)
Grade 1 Free Lunch	-3.735 (2.356)	-7.166 (2.285)***	-3.375 (2.459)	-0.242 (2.060)	Not included in specification
Kindergarten Free Lunch	2.195 (2.073)	1.978 (2.084)	2.116 (1.973)	3.693 (1.968)*	Not included in specification

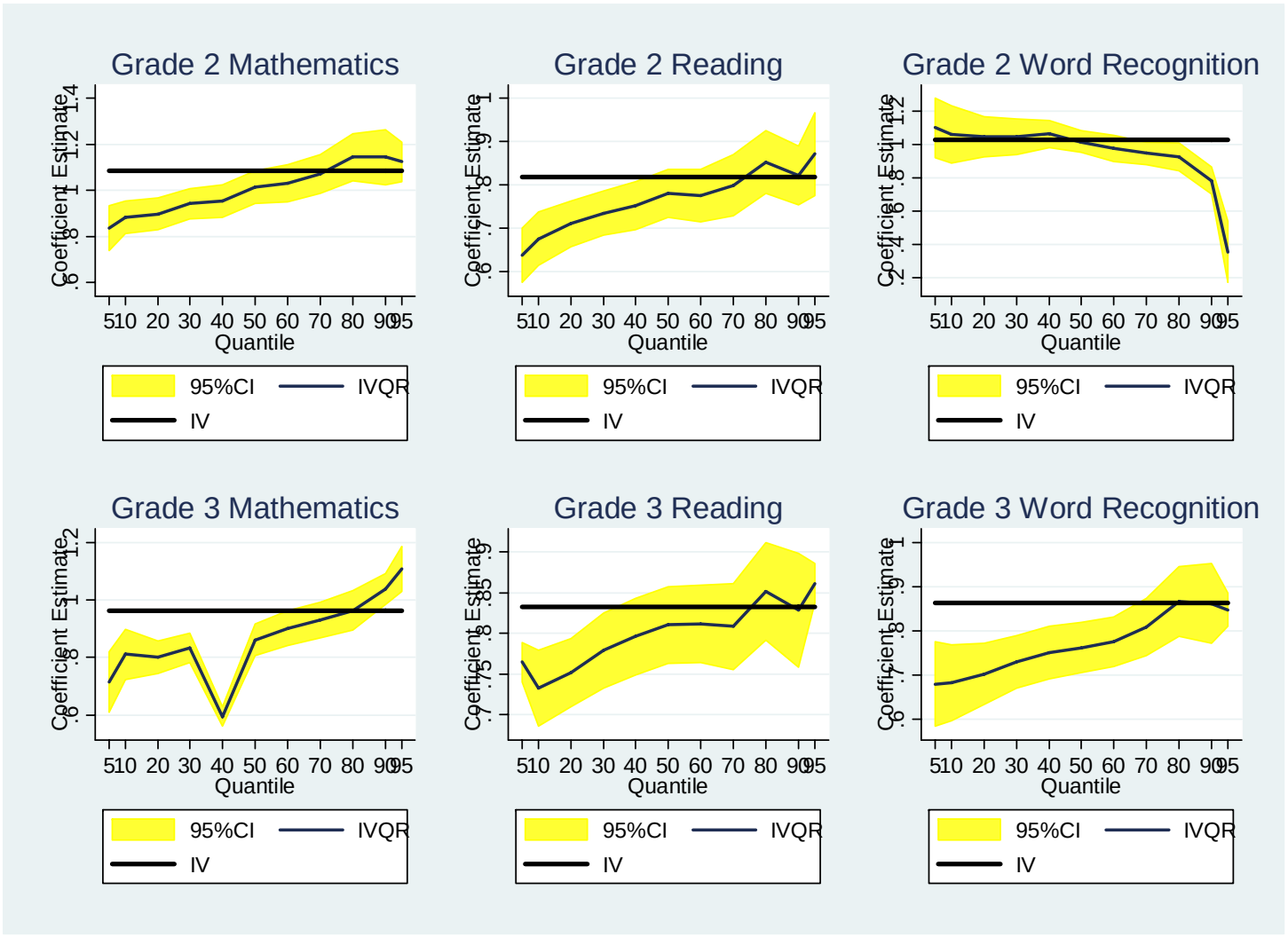
Listening Skills					
Grade 2					
Unobserved Ability Ratio	1.005 (0.036)***	0.683 (0.015)***	β_{iT}=1 is assumed	β_{iT}=0 is assumed	0.997 (0.033)***
Current Class Size	-0.225 (0.226)	-0.314 (0.215)	-0.226 (0.327)	-0.502 (0.347)	-0.198 (0.227)
Female Student	-0.459 (0.924)	-0.451 (0.853)	-0.459 (0.890)	-0.433 (1.218)	-0.491 (0.919)
Student is White/Asian	2.818 (1.944)	6.305 (1.969)***	2.875 (1.940)	13.686 (2.712)***	2.720 (2.140)
Current Free Lunch Status	-2.880 (1.767)	-5.169 (1.545)***	-2.917 (1.571)*	-10.014 (1.944)***	-2.641 (1.612)
Grade 1 Class Size	0.268 (0.262)	0.169 (0.252)	0.266 (0.375)	-0.040 (0.386)	0.377 (0.247)
Kindergarten Class Size	0.189 (0.202)	0.120 (0.192)	0.188 (0.223)	-0.026 (0.245)	Not included in specification
Grade 1 Free Lunch	-0.544 (1.741)	-3.477 (1.633)**	-0.593 (1.725)	-9.684 (2.301)***	-0.248 (1.642)
Kindergarten Free Lunch	0.936 (1.555)	-0.287 (1.483)	0.916 (1.587)	-2.877 (1.809)	Not included in specification
Grade 3					
Unobserved Ability Ratio	0.906 (0.029)***	0.630 (0.016)***	β_{iT}=1 is assumed	β_{iT}=0 is assumed	0.890 (0.030)***
Current Class Size	0.212 (0.254)	0.038 (0.237)	0.219 (0.380)	0.074 (0.446)	-0.037 (0.247)
Female Student	-0.938 (0.926)	-1.371 (0.919)	-0.515 (0.869)	-1.944 (1.369)	-0.992 (0.978)
Student is White/Asian	-0.625 (2.004)	1.613 (2.220)	-2.414 (1.975)	8.487 (3.239)***	-1.271 (2.367)
Current Free Lunch Status	-0.796 (1.739)	-2.030 (1.810)	-0.199 (1.671)	-5.546 (2.353)**	-0.845 (1.849)
Grade 2 Class Size	0.123 (0.317)	0.073 (0.293)	0.166 (0.457)	-0.379 (0.477)	0.182 (0.249)
Grade 1 Class Size	-0.112 (0.317)	-0.001 (0.286)	-0.065 (0.392)	0.169 (0.416)	Not included in specification
Kindergarten Class Size	-0.171 (0.206)	-0.187 (0.207)	-0.178 (0.223)	-0.301 (0.274)	Not included in specification
Grade 2 Free Lunch	0.129 (1.781)	-1.383 (1.917)	1.238 (1.704)	-4.965 (2.379)**	-0.648 (1.837)
Grade 1 Free Lunch	-0.089 (1.757)	-2.860 (1.824)	0.052 (1.774)	-8.242 (2.590)**	Not included in specification
Kindergarten Free Lunch	-0.154 (1.627)	-0.396 (1.663)	-0.613 (1.630)	0.190 (2.006)	Not included in specification

Reading					
Grade 2					
Unobserved Ability Ratio	0.809 (0.025)***	0.607 (0.012)***	β_{iT}=1 is assumed	β_{iT}=0 is assumed	0.819 (0.022)***
Current Class Size	-0.376 (0.283)	-0.383 (0.264)	-0.370 (0.500)	-0.405 (0.501)	-0.386 (0.273)
Female Student	1.692 (1.120)	3.438 (1.051)***	0.044 (1.181)	8.687 (1.607)***	1.548 (1.120)
Student is White/Asian	1.321 (2.332)	2.374 (2.433)	0.329 (2.753)	5.535 (3.233)*	1.462 (2.565)
Current Free Lunch Status	0.109 (2.037)	-2.668 (1.897)	2.729 (2.300)	-11.012 (2.560)***	-0.843 (1.939)
Grade 1 Class Size	0.555 (0.317)**	0.380 (0.309)	0.721 (0.543)	-0.147 (0.525)	0.662 (0.300)**
Kindergarten Class Size	0.102 (0.243)	-0.008 (0.234)	0.204 (0.269)	-0.336 (0.328)	Not included in specification
Grade 1 Free Lunch	1.295 (2.125)	-2.030 (2.006)	4.044 (2.308)*	-11.420 (3.080)***	-0.333 (1.965)
Kindergarten Free Lunch	-3.524 (1.873)**	-4.215 (1.813)**	-2.872 (2.104)	-6.292 (2.702)**	Not included in specification
Grade 3					
Unobserved Ability Ratio	0.841 (0.023)***	0.642 (0.014)***	β_{iT}=1 is assumed	β_{iT}=0 is assumed	0.839 (0.021)***
Current Class Size	-0.578 (0.288)**	-0.421 (0.272)	-0.702 (0.382)*	0.082 (0.462)	-0.632 (0.278)**
Female Student	1.383 (1.106)	2.896 (1.062)***	0.180 (1.274)	7.761 (1.547)***	1.442 (1.115)
Student is White/Asian	2.217 (2.573)	3.351 (2.562)	1.315 (2.665)	6.998 (3.089)**	1.995 (2.661)
Current Free Lunch Status	-0.449 (2.015)	-1.720 (2.087)	0.563 (2.157)	-5.811 (2.672)**	0.624 (2.080)
Grade 2 Class Size	0.769 (0.355)**	0.603 (0.335)	0.901 (0.524)*	0.068 (0.552)	0.488 (0.281)
Grade 1 Class Size	-0.194 (0.362)	-0.307 (0.327)	-0.104 (0.501)	-0.670 (0.521)	Not included in specification
Kindergarten Class Size	-0.274 (0.251)	-0.315 (0.237)	-0.242 (0.290)	-0.445 (0.316)	Not included in specification
Grade 1 Free Lunch	-1.904 (2.197)	-3.322 (2.200)	-0.776 (2.274)	-7.882 (3.164)**	-0.504 (2.069)
Grade 1 Free Lunch	-0.242 (2.060)	-2.170 (2.098)	1.291 (2.401)	-8.371 (3.262)**	Not included in specification
Kindergarten Free Lunch	3.693 (1.968)*	3.384 (1.914)*	3.938 (2.148)*	2.393 (2.691)	Not included in specification

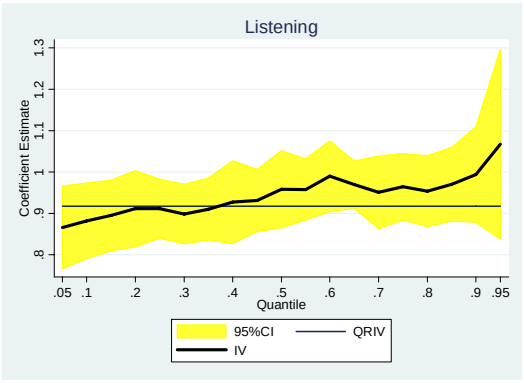
Word Recognition					
Grade 2					
Unobserved Ability Ratio	1.022 (0.033)***	0.637 (0.016)***	β_{iT}=1 is assumed	β_{iT}=0 is assumed	1.023 (0.037)***
Current Class Size	-0.974 (0.402)**	-0.979 (0.360)***	-0.931 (0.540)*	-1.064 (0.612)*	-0.998 (0.390)**
Female Student	-1.936 (1.576)	0.997 (1.415)	-1.920 (1.575)	6.126 (1.949)***	-2.150 (1.587)
Student is White/Asian	0.979 (3.700)	1.647 (3.336)	0.995 (3.633)	2.793 (4.171)	1.315 (3.677)
Current Free Lunch Status	-3.222 (2.747)	-7.719 (2.573)***	-3.523 (2.478)	-15.097 (3.325)***	-3.806 (2.744)
Grade 1 Class Size	1.606 (0.464)***	1.116 (0.423)***	1.576 (0.598)**	0.307 (0.647)	1.534 (0.428)***
Kindergarten Class Size	-0.136 (0.346)	-0.152 (0.322)	-0.171 (0.364)	-0.119 (0.374)	Not included in specification
Grade 1 Free Lunch	2.642 (2.889)	-3.341 (2.707)	2.278 (2.867)	-13.222 (3.933)***	1.751 (2.796)
Kindergarten Free Lunch	-2.221 (2.762)	-2.572 (2.466)	-2.166 (2.700)	-3.285 (3.435)	Not included in specification
Grade 3					
Unobserved Ability Ratio	0.846 (0.026)***	0.549 (0.018)***	β_{iT}=1 is assumed	β_{iT}=0 is assumed	0.841 (0.028)***
Current Class Size	-0.629 (0.414)	-0.308 (0.387)	-0.795 (0.446)*	0.285 (0.536)	-0.644 (0.404)
Female Student	5.496 (1.631)***	6.475 (1.515)***	4.989 (1.751)***	8.281 (1.859)***	5.321 (1.611)***
Student is White/Asian	6.394 (3.905)	5.969 (3.646)	6.615 (4.277)	5.183 (3.954)	6.153 (3.845)
Current Free Lunch Status	1.531 (2.923)	-1.528 (2.949)	3.114 (2.946)	-7.171 (3.359)**	2.869 (3.002)
Grade 2 Class Size	0.962 (0.501)*	0.468 (0.478)	1.217 (0.580)**	-0.443 (0.662)	0.509 (0.408)
Grade 1 Class Size	-0.217 (0.499)	-0.240 (0.467)	-0.205 (0.592)	-0.283 (0.617)	Not included in specification
Kindergarten Class Size	-0.426 (0.346)	-0.456 (0.339)	-0.410 (0.380)	-0.513 (0.410)	Not included in specification
Grade 2 Free Lunch	-1.870 (3.172)	-4.482 (3.150)	-0.518 (3.364)	-9.301 (3.849)**	0.651 (2.986)
Grade 1 Free Lunch	3.008 (2.943)	-0.864 (3.014)	5.012 (3.286)	-8.009 (3.707)**	Not included in specification
Kindergarten Free Lunch	1.707 (3.040)	2.365 (2.731)	1.367 (3.503)	3.578 (3.278)	Not included in specification

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 includes the teacher characteristics. The columns differs in the number of periods of lagged inputs that is listed in the first row and whether the impact of unobserved heterogeneity is fixed and is allowed to be correlated with the inputs. ***,**,* indicate statistical significance at the 1%, 5%, and 10% level respectively. The instruments for each column correspond to IV set 2 in Table 2.

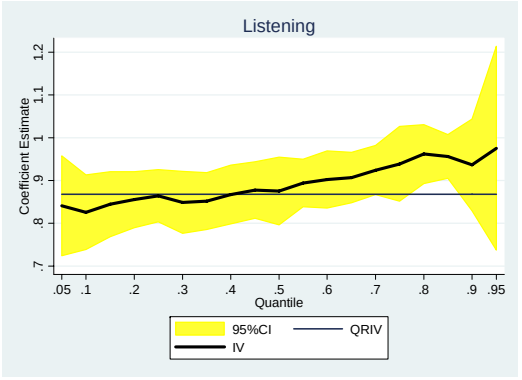
Figure 1: Instrument Variable and Quantile Regression Instrumental Variable Estimates of the Impacts of Unobserved Ability



GRADE 2 Listening



GRADE 3 Listening



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: Results from Model Specification Tests of the Education Production Function

Outcome→ Alternative Model ↓	Grade 2 Mathematics Test Score	Grade 2 Reading Test Score	Grade 2 Word Recognition Test Score	Grade 2 Listening Skills Test Score
Equation 7 where $\beta_{iT}=0$ is assumed	566.94 (1.000)	6.65 (0.010)	69.30 (0.000)	482.82 (1.000)
Equation 7 where $\beta_{iT}=1$ is assumed	841.60 (0.000)	978.39 (0.000)	680.08 (0.000)	713.64 (0.000)
Equation 8 Contemporaneous Model	883.69 (0.000)	1022.50 (0.000)	710.61 (0.000)	760.88 (0.000)
Equation 9 Value Added Model	784.98 (0.000)	951.51 (0.000)	499.70 (0.000)	898.77 (0.000)
Equation 10 Linear Growth Model	543.56 (0.000)	36.80 (0.000)	32.05 (0.000)	456.42 (0.000)
Outcome→ Alternative Model ↓	Grade 3 Mathematics Test Score	Grade 3 Reading Test Score	Grade 3 Word Recognition Test Score	Grade 3Listening Skills Test Score
Equation 7 where $\beta_{iT}=0$ is assumed	21.81 (0.000)	251.34 (0.000)	754.29 (0.000)	76.29 (0.000)
Equation 7 where $\beta_{iT}=1$ is assumed	918.99 (0.000)	1153.62 (0.000)	1044.22 (0.000)	668.44 (0.000)
Equation 8 Contemporaneous Model	964.50 (0.000)	1193.61 (0.000)	1075.71 (0.000)	719.21 (0.000)
Equation 9 Value Added Model	388.89 (0.000)	294.29 (0.000)	129.50 (0.000)	567.82 (0.000)
Equation 10 Linear Growth Model	52.06 (0.000)	303.61 (0.000)	803.06 (0.000)	51.19 (0.000)

Note: Each entry contains the likelihood ratio test statistic and associated –value that compares the instrumental variables estimates of equation 7 with IV set 2 against an alternative specification listed in the row variable for each grade subject area presented by columns. All of the entries are significantly different at the 1% level or lower.

Appendix Table 2: Tests of Serial Correlation in the Residuals to Validate Using Lagged Test Scores as Instrumental Variables.

Outcome Equation → Instrument in First Stage Equation ↓	Grade 2 Mathematics Test Score	Grade 2 Reading Test Score	Grade 2 Word Recognition Test Score	Grade 2 Listening Skills Test Score
Kindergarten Mathematics	Not included as an instrument	2.092** (0.036)	1.280 (0.201)	-0.186 (0.852)
Kindergarten Reading	0.233 (0.815)	Not included as an instrument	-0.237 (0.813)	0.820 (0.412)
Kindergarten Word Recognition	0.216 (0.829)	-0.904 (0.366)	Not included as an instrument	0.793 (0.428)
Kindergarten Listening Skills	-0.034 (0.973)	1.422 (0.154)	2.082** (0.037)	Not included as an instrument
Outcome Equation → Instrument in First Stage Equation ↓	Grade 3Mathematics Test Score	Grade3 Reading Test Score	Grade 3 Word Recognition Test Score	Grade 3 Listening Skills Test Score
Kindergarten Mathematics	Not included as an instrument	1.109 (0.267)	1.347 (0.178)	0.960 (0.337)
Kindergarten Reading	0.606 (0.544)	Not included as an instrument	1.144 (0.253)	0.098 (0.922)
Kindergarten Word Recognition	1.277 (0.201)	-0.015 (0.988)	Not included as an instrument	0.701 (0.483)
Kindergarten Listening Skills	-0.411 (0.681)	1.526 (0.127)	-0.400 (0.689)	Not included as an instrument
Grade 1 Mathematics	Not included as an instrument	1.109 (0.267)	1.347 (0.178)	0.959 (0.337)
Grade 1 Reading	-0.212 (0.832)	Not included as an instrument	-1.333 (0.183)	-0.438 (0.663)
Grade 1 Word Recognition	7.158*** (0.000)	-1.104 (0.270)	Not included as an instrument	8.589*** (0.000)
Grade 1 Listening Skills	-0.411 (0.681)	1.261 (0.063)	0.400 (0.689)	Not included as an instrument

Note: Each cell contains the test statistic that is distributed standard Normal and the p-value of the test that the residuals in the row column pair are uncorrelated. . ***,**,* indicate statistical significance at the 1%, 5%, and 10% level respectively.

Appendix Table 3: Hausman Test Results Comparing OLS and IV Estimates of Equation 7

Subject Area	Mathematics	Reading	Word Recognition	Listening Skills
Full Specification				
Grade 2	470.20 (0.00)***	113.42 (0.00)***	140.43 (0.00)***	92.40 (0.00)***
Grade 3	207.10 (0.00)***	187.83 (0.00)***	236.69 (0.00)***	116.98 (0.00)***
Only the Coefficient on the Unobserved Ability Ratio Term				
Grade 2	13.56 (0.00)***	10.665 (0.00)***	11.85 (0.00)***	9.612 (0.00)***
Grade 3	14.389 (0.00)***	13.705 (0.00)***	15.439 (0.00)***	11.031 (0.00)***

Note: Each cell contains the test statistic and the p-value of a Hausman test where under the Null, the OLS estimator is consistent and efficient. Estimates from columns 1 and 2 of Table 4 with are used to conduct the tests. ***,**,* indicate statistical significance at the 1%, 5%, and 10% level respectively.