The Contribution of Early Childhood and Schools to Cognitive Gaps: New Evidence from Peru

Juan F. Castro[1]
Universidad del Pacifico

Caine Rolleston
University College London

Abstract

Cognitive gaps between children of different socioeconomic backgrounds are particularly significant in the developing world. We propose and use a new decomposition strategy to measure the contribution of early childhood and school influences to the cognitive gap between urban and rural eight-year-old children in Peru. This strategy accounts for the relation between family choices and skill inputs and is less prone to biases than those employed before. We find that school influences occurring between ages 6 and 8, account for a significant share of urban/rural cognitive gap (around 35%). The share attributable to early childhood influences is important but no larger than 50%. Because skill depreciates, only a fraction of the gap (70-80%) is carried forward to the next period. Therefore, inequalities in school environments are sustaining a cognitive gap that would otherwise be smaller and this explains why differences that emerge during early childhood can remain unchanged after children start school.

JEL codes: I24, O15, C18
Keywords: cognitive gap; linear decomposition; early childhood, school inputs.

[1] Corresponding author. Department of Economics at Universidad del Pacifico. Av. Salaverry 2020, Lima 11, Peru. Email: castro_jf@up.edu-pe. Phone: 511 219 0100. Fax: 511 219 0135. The authors are grateful to Sandra McNally, two anonymous reviewers, Paul Glewwe, Douglas Gollin, Imane Chaara, Andreas Georgiadis, Stefan Dercon and participants at the 2016 Meeting of the International Association for Applied Econometrics, the 2016 Meeting of Latin American and Caribbean Economic Association and the 12th Midwest International Economic Development Conference for their valuable comments.

1. Introduction and motivation

Differences in developmental outcomes between children of dissimilar socioeconomic backgrounds are particularly significant in the developing world (Grantham-McGregor et al., 2007; Walker et al., 2007). Peru is no exception to the presence of these early forms of inequality. In fact, national student evaluations reveal a significant and persistent gap between the proportion of urban and rural second grade students that attain satisfactory results in reading comprehension (see Figure 1).[2] Evidence on cognitive skill formation confirms the presence of significant developmental disparities between urban and rural school-age children (see Figure 2).

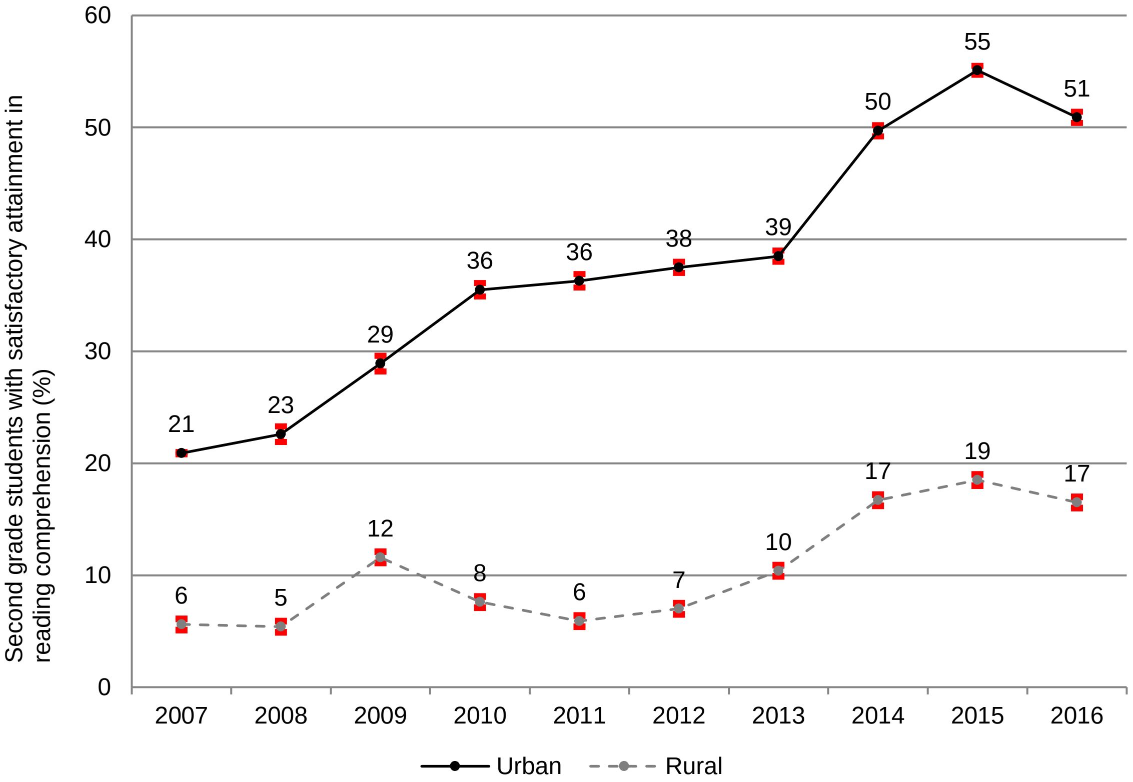
Cognitive skill formation is a cumulative process and, thus, all relevant influences that had been exerted before a skill is measured can play a role in shaping these gaps. A relevant question that follows concerns which particular influence or group of influences play a significant role for the emergence of these differences. Do earlier influences matter more than those occurring later in the life of children? Do influences originating in a particular environment (such as these children's home or school) play a major part? This study assesses the importance of early childhood and school influences for the cognitive skill gap observed between urban and rural eight-year-old children in Peru.

Longitudinal evidence available for developed and developing countries shows that differences in cognitive skill between advantaged and disadvantaged children emerge before children start school and remain fairly unchanged throughout their school years (Heckman, 2006, 2007; Paxson and Schady, 2007; Schady et al., 2014). This fact, combined with the

[2] The gap was around 30 percentage points between 2010 and 2013, and has increased in recent years due to a faster improvement in urban areas. This correlates with increased access to learning materials and better infrastructure in public urban schools but there are numerous potential reasons behind this trend, including an improvement in early childhood environments. A decomposition exercise similar to the one proposed here but applied to the evolution of cognitive test scores could be used to learn more about this matter.

observation that difference in school characteristics usually favor advantaged children, has been put forward by some authors as evidence that schools contribute little to cognitive development, so investment efforts aimed at closing cognitive skill gaps should concentrate during early childhood (Cunha et al., 2006; Heckman, 2006; Schady et al., 2014). There is, however, an implicit "difference-in-differences" approach in this interpretation which relies on the existence of perfect persistence (or no depreciation) in skill formation.[3]

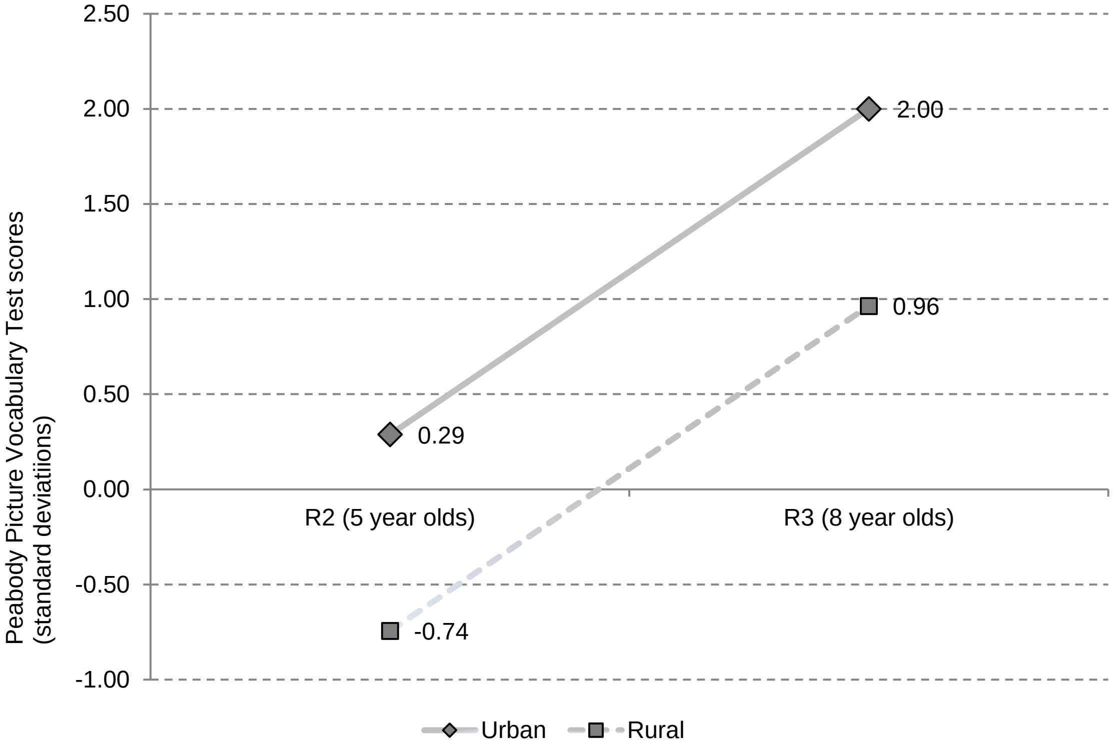
Figure 1
Peru: proportion of second grade students with a satisfactory attainment in reading comprehension (urban and rural point estimates and 95% confidence intervals)



Source: Control Sample of the National Student Evaluation (2007-2016), Ministry of Education - Peru.

[3] To see this, one can refer to Figure 2 and consider the difference in school environments between urban and rural children as a treatment applied to rural children between ages 5 and 8. Given the cumulative nature of skill, judging the importance of this treatment by subtracting the skill gap measured at age 5 from the skill gap measured at age 8 assumes that the entire age-5 gap persists up to age 8. If skill is subject to depreciation, this assumption will not hold. Notice that this is similar to the common trends assumption required for a difference-in-differences estimation. If only a fraction of skill is carried forward from one period to the other and there are no further inequalities, the gap will grow smaller and the evolution of skill of urban and rural children will not be parallel.

Figure 2
Peru: urban/rural differences in cognitive achievement (standardized Peabody Picture Vocabulary Test scores)



	Round 2	Round 3	Obs.
Urban	0.29	2.00	1,276
Rural	-0.74	0.96	493
Gap	1.03***	1.04***	

Significant at the 1% (***), 5% (**), 10% (*).
Source: Own calculations based on the Peruvian Young Lives database (Younger Cohort; rounds 2 and 3).

Another body of evidence regarding the importance of schools for cognitive skill formation and educational outcomes is provided by those studies that have performed linear decompositions of learning outcome gaps to measure the contribution of school and home influences. Decomposition exercises available for several developing countries seem to corroborate that schools play only a subsidiary role as they have found that household characteristics make a larger contribution than schools to developmental gaps (Arteaga and Glewwe, 2014; Hernandez-Zavala et al., 2006; McEwan and Marshall, 2004; Ramos et al., 2012). Peru is among the countries that exhibit this kind of evidence.

Most of the studies in this strand of the literature have relied on some form of Blinder-Oaxaca decomposition (Blinder, 1973; Oaxaca, 1973). There are several ways to implement this technique so one needs to choose which specific strategy to follow and how to interpret its components. In addition, one has to devise a rule to classify the contribution of individual variables into different categories (e.g. “home” or “school” influences). In this regard, a revision of the studies applied to the developing world reveals that the rule commonly employed to separate home and school influences has been to assign all observed household, family and child characteristics to the home influences category. This is problematic because some of these characteristics can have indirect effects that operate through both home and school environments.

A good example of this is family income or wealth. More affluent families can provide a more nurturing environment to their children at home but can also purchase better quality school inputs. Therefore, grouping all household, family and child characteristics (including family income) under the “home influences” category entails the risk of overstating the importance of these influences. The underlying assumptions if one follows this rule are that all relevant school influences have been accounted for or that school influences are unrelated to families’ choices. These are both strong assumptions, especially if information on school and teacher characteristics is scarce or the analysis is carried out for a heterogeneous schooling system. This potential source of bias has been overlooked so far in the literature and can affect any decomposition exercise that relies on an empirical specification which includes variables reflecting household, family or child characteristics as controls.[4]

[4] Other studies within the same strand of the literature have relied on school fixed effects to account for the contribution of school influences and have found that these make a significant contribution to cognitive gaps (McEwan, 2004; McEwan and Trowbridge, 2007). The risk in this case is of overstating the importance of schools. School fixed effects absorb all direct influences that are invariant within schools and school inputs are among these. The

This paper contributes to the literature in two ways. First, it provides new evidence about the importance of early childhood and school influences for cognitive gaps in a developing country setting. Second, it presents and uses a decomposition strategy less prone to biases than those employed thus far.

We use the insights of a model that describes the production of skill and how families' choices influence this process to distinguish between skill inputs (those influences that have a direct effect on skill and belong to its production function, such as the availability of books at home) and input determinants (those that determine families' decisions and belong to the demand function of skill, such as family wealth). Instead of grouping all child, family and household characteristics together, those classified as input determinants are assigned to a special category hosting the contribution of omitted inputs in general (i.e. inputs that could belong to the early childhood, school or home environment). This is because these characteristics can affect skill through inputs that belong to any of these environments. This "omitted inputs" category resembles the "unexplained" component in a standard Blinder-Oaxaca decomposition and its use prevents us from making strong assumptions about the source of unobservable inputs.

We find that school influences occurring between ages 6 and 8, account for a significant share of urban/rural cognitive gap (around 35%). The share attributable to early childhood influences is important but no larger than 50%. In terms of the evidence presented in Figure 2, these results imply that the gap observed at age 8 would be around 35% smaller if urban and rural schools offer similar environments.

assumption required for this strategy to identify the contribution of "school influences" is that these influences are the only inputs shared by students that belong to the same school. This is hardly the case in a context where there is strong correlation between children's socioeconomic status and the quality of schooling received.

The gap would be smaller because only a fraction of the difference is carried forward to the next period. In fact, we found evidence of less than perfect persistence in skill formation: between 34 and 52% of the gap observed at age 5 persists up to age 8. This means that the difference in cognitive skill at age 8 is similar to that observed at age 5 not because schools do not matter but because inequalities in school environments are sustaining a gap that would otherwise be smaller.

The rest of the paper is organized as follows. Section 2 presents a framework describing the skill formation process and how families' choices determine its inputs, allowing for endogenous school quality. In section 3, we use the insights of this model to propose a decomposition strategy and further explain why it is less prone to biases than those employed thus far. We also explain how this strategy relates to the standard Blinder-Oaxaca technique. In section 4, we apply this decomposition strategy to measure the contribution of early childhood and school influences to the cognitive gap observed between urban and rural 8-year-old children in Peru. We also compare the results of our decomposition strategy against those of the standard Blinder-Oaxaca technique employed in the literature to illustrate how the latter can produce biased results. Section 5 closes with some final remarks.

2. The production function of skill and families' choices

In this section, we describe the skill formation technology and present a simple model describing how families' choices determine its inputs. For this, let us divide the relevant phase of child development into two time periods. The first begins when the child is born and finishes at age 5, that is, when the child is ready to start the basic education cycle. The second period corresponds to the time when the child remains within primary school age, which is usually between ages 6 and 11.

Assume that skill demonstrated by child i at the end of period 2 (A_{i2}) is a function of contemporaneous and past direct influences affecting the child. This is consistent with the notion that skill formation is a cumulative process. Formally:

$$A_{i2} = A_2(H_{i2}, H_{i1}, S_{i2}, SY_{i2}, h_{i2}, h_{i1}, f_i, \mu_{i0}) \qquad (1)$$

where H_{i1} are the educational inputs provided during early childhood (period 1); H_{i2} are the educational inputs provided at home during period 2; S_{i2} are the educational inputs provided at the school where the child is enrolled during period 2; SY_{i2} are the years of schooling attained during period 2; h_{it} indicates the child's health status during period t; f_i captures predetermined direct influences; and μ_{i0} is the child's innate ability.

Importantly, expression (1) denotes a structural relationship between skill and those variables that have a direct effect on it. These variables will reflect the environment surrounding the child (characterizing activities, materials and individuals), as well as child characteristics that influence directly the acquisition of skill. As stressed in Glewwe and Miguel (2008), all the variables in the production function should affect skill directly, and all the variables with a direct effect should be included in this function. We further classify these direct influences as inputs (if they are determined by families' choices during the period under analysis) or as predetermined (if they are outside the current choice set of families).

The rest of the model follows Glewwe and Miguel (2008) closely but extends their original formulation to allow for endogenous school inputs and differences in the supply of educational services available to each family. Glewwe and Miguel (2008) assume that school and teacher characteristics available to the child are not influenced by parental decisions made during the period under analysis. During this period, families' choices related to the school environment are limited to the number of years of schooling.

It is reasonable to assume, especially in highly unequal schooling systems, that parents can influence the school and teacher characteristics available to their children either by changing location or because localities are characterized by a distribution of educational services from where parents can choose. In this setting, the simplest assumption is that all families can choose a school from a common pool or choice set (see, for example, Todd and Wolpin (2003)). This is consistent with a situation where there is a similar distribution of schooling services across localities or migration costs are not significant.

In what follows, we will adopt a more flexible approach. The model will predict that families choose a number of schooling years at a school (j) with a particular set of characteristics (S_{ij}) for a given price (p_{ij}). We assume that school characteristics are chosen from a given set $Q_i = \{S_{i1}, \ldots, S_{iJ_i}\}$ and that this set is not necessarily the same for all families. This set is defined by the distribution of educational services available in the geographical area within which the family makes its location decisions.

Consistent with the two-period setting assumed above, consider that parents maximize the following utility function:

$$U_i = U(C_{i1}, C_{i2}, h_{i2}, h_{i1}, A_{i2}; \tau, \sigma, \omega) \tag{2}$$

where C_{it} is child i's parental consumption of an aggregate good in period t, and τ, σ and ω reflect parental preferences regarding time, child's skill and child's health, respectively.

Child health is determined according to the following production functions:

$$h_{i1} = G_1(c_{i1}, M_{i1}, HE_i, \eta_{i0}) \tag{3}$$

$$h_{i2} = G_2(h_{i1}, c_{i2}, M_{i2}, HE_i, \eta_{i0}) \tag{4}$$

where c_{it} is child i's consumption of the aggregate good in period t, M_{it} are health inputs provided in period t, HE_i captures the local health environment, and η_{i0} is the child's innate healthiness.

In this setting, parents choose consumption levels (C_{it} and c_{it}), health inputs (M_{it}), educational inputs provided during early childhood and at home (H_{i1}, H_{i2}), and years of schooling at the different schools available to them $\left(SY_{ij}; j = 1, \dots, J_i\right)$ to maximize utility given in (2). This is done subject to the skill formation technology given in (1), the production functions for health given in (3) and (4), and the following budget constraint:

$$Y_{i1} - V_{i1} = p_{c1}(C_{i1} + c_{i1}) + p_{m1}M_{i1} + p_{h1}H_{i1} \tag{5}$$

$$Y_{i2} + (1 + r)V_{i1} = p_{c2}(C_{i2} + c_{i2}) + p_{m2}M_{i2} + p_{h2}H_{i2} + \sum_{j=1}^{J_i} p_{ij}SY_{ij} \tag{6}$$

In (5) and (6), V_{i1} represent savings, p_{ct} is the price of the aggregate consumption good in period t, p_{mt} is the price of health inputs in period t $(t = 1,2)$, p_{h1} is the price of educational inputs provided during early childhood, p_{h2} is the price of educational inputs provided at home during period 2, p_{ij} is the price of one year of schooling at school with characteristics S_{ij}, and SY_{ij} is the number of years of schooling demanded from school j. Y_{it} is period t exogenously determined income, and r is the interest rate at which parents are assumed can borrow or lend between the two time periods.

As already explained, we assume that parents will choose a particular school from a given set. We also assume that there is no school switching. This means that we need an additional set of restrictions to fully characterize the optimization problem faced by parents. Formally:

$$SY_{i2} = \max_j\left(SY_{ij}\right) \tag{7}$$

$$S_{i2} = S_{ij} \; if \; SY_{ij} = SY_{i2}; \; j = 1, \dots, J_i \tag{8}$$

Condition (7) guarantees that families will choose a single school. Condition (8) matches the school inputs relevant for the production of skill to the characteristics of the chosen school.

The first order conditions of the problem stated above provide the relationships explaining the optimal levels of consumption, health inputs, educational home inputs, years of schooling and school inputs. All of these demand functions depend on: (i) resources (Y_{i1}, Y_{i2}); (ii) prices (r, p, p_{ij}) $j = 1, \dots, J_i$ and $p = (p_{c1}, p_{c2}, p_{m1}, p_{m2}, p_{h1}, p_{h2})$; (iii) exogenous environmental variables (HE_i, Q_i); (iv) predetermined direct influences (f_i); (v) endowments (μ_{i0}, η_{i0}); and (vi) preferences (τ, σ, ω). For example, the demand functions for the educational inputs of cognitive skill can be expressed as:

$$H_{it}^{*} = H_t\left(Y_{i1}, Y_{i2}; r, p, p_{i1}, \dots, p_{iJ_i}; HE_i, Q_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega\right) \; t = 1,2 \tag{9}$$

$$SY_{i2}^{*} = \max_j\left(SY_{ij}^{*}\right) = \max_j\left[SY_j\left(Y_{i1}, Y_{i2}; r, p, p_{i1}, \dots, p_{iJ_i}; HE_i, Q_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega\right)\right] \tag{10}$$

$$S_{i2}^{*} = S_{ij} \; if \; SY_{ij}^{*} = SY_{i2}^{*}; \; j = 1, \dots, J_i \tag{11}$$

These functions describe how differences in the availability of school services (captured through differences in Q_i) can lead to differences in the educational inputs provided to the child during early childhood and also, later, at school. It also describes how families that share the same Q_i can end up offering different school inputs to their children, depending on variables such as their income.

The production function indicated in (1) involves only and all of the variables that have a direct effect on skill, whether they are predetermined or not. In addition to this function, there are three other meaningful relations that can be postulated to explain children’s skill: a demand function, a conditional demand function, and a hybrid production function. We briefly describe them below to clarify what is meant by a hybrid production function as it will play an important role in the estimations that follow.[5]

The demand function for cognitive skill involves only predetermined variables that can have a direct or indirect effect on skill. It can be obtained by replacing all the inputs of skill in the production function by their corresponding demand function. This yields:

$$A_{i2} = A_2^D\left(Y_{i1}, Y_{i2}; r, p, p_{i1}, \dots, p_{iJ_i}; HE_i, Q_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega\right) \tag{12}$$

The conditional skill demand function conditioned over input k, only involves input k and the exogenous determinants of the rest of inputs. To obtain this relation we need first to consider the conditional demand functions for the rest of inputs, conditioned over input k. These are obtained by fixing input k at its utility maximising level, which implies that prices related to input k and resources devoted to its consumption are no longer relevant arguments of the demand for the rest of inputs.

For example, conditional demand functions for educational inputs provided during early childhood and at home, conditioned over school inputs (years of schooling and school characteristics) are given by:

$$H_{it}^{CD} = H_t(S_{i2}, SY_{i2}; Y_{CD};\ p; HE_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega)\ \ t = 1{,}2 \tag{13}$$

where $Y_{CD} =\ Y_{i1} + \frac{Y_{i2}}{1+r} - \frac{\sum_{j=1}^{J_i} p_{ij} SY_{ij}}{1+r}$ refers to resources after adjusting for school expenditures. As already noted, the price of schooling $\left(p_{ij}\right)$ is no longer present in (13).

Similar expressions can be obtained for the conditional demand for child’s health in both periods after building conditional demand functions for child’s consumption and health

[5] For a complete description of how to obtain and interpret a demand and a conditional demand function, the reader can consult Glewwe and Miguel (2008).

inputs. Replacing conditional demand functions for early childhood and educational home inputs and child's health in the production function given in (1) yields the demand for child's skill conditioned over school inputs. Formally:

$$A_{i2} = A_2^{CD}(S_{i2}, SY_{i2}; Y_{CD}; p; HE_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega) \qquad (14)$$

Finally, and following the example centered on school inputs, a hybrid production function can be obtained if we replace all inputs in (1), except those related to the school environment, by their respective demand functions. Doing this we obtain:

$$A_{i2} = A_2^{H}\left(S_{i2}, SY_{i2}; Y_{i1}, Y_{i2}; r, p, p_{i1}, \dots, p_{iJ_i}; HE_i, Q_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega\right) \qquad (15)$$

The motivation for this type of specification is empirical as it aims at recovering the production function parameters of observed inputs evading omitted variable biases by replacing unobserved inputs with their corresponding demand functions (Rosenzweig and Schultz, 1983; Todd and Wolpin, 2007).

There is an important difference between the effect of school inputs provided by equations (14) and (15). Just like in the production function given in (1), the effect of school inputs captured in the hybrid function in equation (15) corresponds to the direct effect of these inputs on skill, holding all other direct influences constant. The effect of school inputs provided by the conditional demand function in expression (14) includes this direct effect but also captures the indirect effect produced through changes in other inputs. These changes can occur because families respond to the initial shock in one of the inputs. Experimental designs and instrumental variable techniques will typically identify the parameters of a conditional demand function (or the "policy effects" as denoted in Todd and Wolpin (2003)).[6]

[6] Notice that in an experimental setting, post-treatment values of other inputs can change in response to introducing exogenous variation in a certain input.

3. Empirical specifications and decomposition strategy

In this section, we use the insights provided by the model described above to propose a decomposition strategy. We explain the assumptions required for the identification of the contributions of early childhood and school influences and explain why this strategy is less prone to biases than those employed thus far in the literature. We also discuss its rationale under the lens of the Blinder-Oaxaca (henceforth BO) technique. The decomposition strategy proposed here is motivated by the empirical goal of decomposing an urban/rural gap in cognitive skill. The main messages of the analysis, however, can be generalized to situations that involve gaps observed between other groups of children.

3.1. Cumulative and value added specifications

Let us assume that the production function given in (1) is approximately linear. This allows one to express the production function of skill as follows:[7]

$$A_{i2} = H'_{i2}\gamma_1 + H'_{i1}\gamma_2 + S'_{i2}\phi_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f_i{}'\lambda_2 + \mu_{i0}\beta_2 \qquad (16)$$

This can be described as a "cumulative" model where skill demonstrated at the end of period 2 is expressed as a function of all relevant direct influences that took place until that moment. It is also possible to express A_{i2} as a function of period 1 skill and period 2 influences only. To see this, consider that period 1 skill can be written as:

$$A_{i1} = H'_{i1}\gamma_1 + h_{i1}\varphi_1 + f_i{}'\lambda + \mu_{i0}\beta \qquad (17)$$

[7] Assume, for simplicity, that the number of years of schooling (SY_{i2} in (1)) is contained in the vector of school inputs S_{i2}.

Notice that parameters λ and β in (17) indicate the contemporaneous effect of predetermined direct influences and innate ability, respectively, while parameters λ_2 and β_2 in (16) express the cumulative effect (until period 2) of this same pair of influences.

If we subtract ρA_{i1} from (16) and assume that the effect of inputs decays at a rate ρ we obtain:

$$A_{i2} = \rho A_{i1} + \ H'_{i2}\gamma_1 + S'_{i2}\phi_1 + h_{i2}\varphi_1 + f_i{}'\lambda + \mu_{i0}\beta \qquad (18)$$

where $\lambda = \lambda_2 - \rho\lambda$ and $\beta\ = \beta_2 - \rho\beta$. The expression given in (18) is known as a “value added” model (Todd and Wolpin, 2003).

As already explained, the objective of a hybrid specification is to control for omitted inputs using the arguments of their corresponding demand function. Following the results of the model of family choice presented in the previous section, the demand functions of the inputs of skill (including those omitted) depend on predetermined household, family and child characteristics that influence skill directly (f_i) and other exogenous input determinants capturing differences in resources, prices, environments and preferences.

In what follows, we will allow the geographical domain to be a potentially relevant argument in the demand function of inputs. In fact, it can control for differences in exogenous environmental variables such as the general health status or the availability of educational services in the area (HE_i, S_i in the model presented above, respectively). For this to be a reasonable assumption, the household’s geographical domain should not be part of the choices made by families during the period under analysis. To ease the exposition that follows, we will consider this potential input determinant separately from the rest. Let the indicator G_i denote the geographical domain ($G_i = 1$ if the child lives in the urban area and

$G_i = 0$ if he or she lives in the rural domain) and assume that the rest of exogenous input determinants are contained in vector z_i.

Let us shift to the empirical versions of (16) and (18) assuming that cognitive skill is measured with error through the scores obtained in some test: $T_{i2} = A_{i2} + \varepsilon_{i2}; E(\varepsilon_{i2}) = 0, Cov(\varepsilon_{i2}, A_{i2}) = 0$. Also assume there can be omitted inputs and that these have been replaced by their corresponding demand functions to obtain:

$$T_{i2} = H'_{i2}\gamma_1 + H'_{i1}\gamma_2 + S'_{i2}\phi_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f_i{}'\pi + z'_i\psi + \theta G_i + e^{H}_{i2} \qquad (19)$$

$$T_{i2} = \rho T_{i1} + H'_{i2}\gamma_1 + S'_{i2}\phi_1 + h_{i2}\varphi_1 + f_i{}'\tilde{\pi} + z'_i\tilde{\psi} + \tilde{\theta} G_i + e^{VA}_{i2} \qquad (20)$$

Equations (19) and (20) are the empirical versions of the cumulative and value added models, respectively. They are also hybrid versions because they include exogenous input determinants z_i and G_i.

3.2. Blinder-Oaxaca with a twist

Let us define the cognitive gap as the difference in expected skill between children belonging to the urban and rural domains: $E(A_{i2}|U) - E(A_{i2}|R)$. Its empirical counterpart is given by: $\bar{T}_{U2} - \bar{T}_{R2}$, where upper bars indicate sample means. The inclusion of the group indicator in (19) and (20) ensures that an OLS regression passes through the mean of both groups. Thus, for the cumulative specification we have:

$$\bar{T}_{U2} - \bar{T}_{R2} = (\bar{H}_{U2} - \bar{H}_{R2})'\hat{\gamma}_1 + (\bar{H}_{U1} - \bar{H}_{R1})'\hat{\gamma}_2 + (\bar{S}_{U2} - \bar{S}_{R2})'\hat{\phi}_1 + \left(\bar{h}_{U2} - \bar{h}_{R2}\right)\hat{\varphi}_1 + \left(\bar{h}_{U1} - \bar{h}_{R1}\right)\hat{\varphi}_2 + \left(\bar{f}_U - \bar{f}_R\right)'\hat{\pi} + (\bar{z}_U - \bar{z}_R)'\hat{\psi} + \hat{\theta} \qquad (21)$$

And, for the value added model we have:

$$\bar{T}_{U2} - \bar{T}_{R2} = (\bar{T}_{U1} - \bar{T}_{R1})\hat{\rho} + (\bar{H}_{U2} - \bar{H}_{R2})'\hat{\gamma}_1 + (\bar{S}_{U2} - \bar{S}_{R2})'\hat{\phi}_1 + \left(\bar{h}_{U2} - \bar{h}_{R2}\right)\hat{\varphi}_1 +$$

$$\left(\bar{f}_U - \bar{f}_R\right)'\hat{\tilde{\pi}} + (\bar{z}_U - \bar{z}_R)'\widehat{\tilde{\psi}} + \hat{\tilde{\theta}} \qquad (22)$$

Recall that the empirical goal is to decompose $\bar{T}_{U2} - \bar{T}_{R2}$ between influences originated during early childhood and influences originated at school. Both specifications allow one to identify the contribution of observed school inputs. The cumulative specification can provide an estimate of the contribution of early childhood influences based on the contribution of period 1 educational and health inputs. The value added model aggregates all early childhood influences into a single component $\left((\bar{T}_{U1} - \bar{T}_{R1})\hat{\rho}\right)$ containing the contribution of early childhood educational and health inputs, and also the contribution due to the early childhood effect of predetermined direct influences and innate ability.

An important decision concerns where to assign the contribution of predetermined direct influences (f_i) and exogenous input determinants (z_i, G_i). The logic behind a hybrid specification indicates that the parameters contained in vectors π and $\tilde{\pi}$ in (19) and (20), respectively, are a function of: (i) the production function parameters of f_i; (ii) the production function parameters of the omitted inputs in the cumulative and value added specifications, respectively; and (iii) the parameters relating these omitted inputs to f_i in their corresponding demand equations. Parameters contained in vectors ψ and θ in (19) are a function of: (i) the production function parameters of the omitted inputs in the cumulative specification; and (ii) the parameters of the demand equations of these omitted inputs. The same is true for parameters in $\tilde{\psi}$ and $\tilde{\theta}$ in (20), but with respect to the omitted inputs in the value added model.

An important implication of this parameter structure is that it will not be possible to separately identify the direct and indirect effects of predetermined direct influences (f_i)

unless we assume there are no omitted inputs or that variables contained in (z_i, G_i) are sufficient to characterize the demand equation of omitted inputs. In addition, if we want to assign the contribution of variables in f_i, z_i and G_i to either the school or early childhood category, we have to assume that there are no omitted inputs from the other category or that omitted inputs from the other category do not depend on families' choices.

The decomposition rule employed in the literature consists in grouping all household, family and child characteristics together. This relies on the assumption that all school inputs have been accounted for or that school inputs are not determined by families' choices. These are both strong assumptions, especially if the information on school characteristics is scarce or the education system exhibits heterogeneous quality and families can choose where to enroll their children. If none of these assumptions hold, the rule employed in the literature will produce biased results. In Appendix 1, we illustrate this for the case of a linear demand function.

To avoid relying on these strong assumptions, the decomposition strategy proposed here assigns the contribution of all variables contained in f_i and z_i and the indicator G_i into a special category hosting the contribution of predetermined direct influences and omitted inputs in general. For a value added specification, this joint contribution will account only for period 2 predetermined direct influences and period 2 omitted inputs. Table 2 summarizes the categories proposed based on the contributions given in (21) and (22).

Table 2
Decomposition categories and variables included in each category

Cumulative specification	
Early childhood educational and health inputs	$(\overline{H}_{U1} - \overline{H}_{R1})'\hat{\gamma}_2 + (\overline{h}_{U1} - \overline{h}_{R1})\hat{\varphi}_2$
School inputs	$(\bar{S}_{U2} - \bar{S}_{R2})'\hat{\phi}_1$
Period 2 home and health inputs	$(\overline{H}_{U2} - \overline{H}_{R2})'\hat{\gamma}_1 + (\overline{h}_{U2} - \overline{h}_{R2})\hat{\varphi}_1$
Predetermined direct influences and omitted inputs	$(\bar{f}_U - \bar{f}_R)'\hat{\pi} + (\bar{z}_U - \bar{z}_R)'\hat{\psi} + \hat{\theta}$
Value added specification	
Early childhood influences	$(\overline{T}_{U1} - \overline{T}_{R1})\hat{\rho}$
School inputs	$(\bar{S}_{U2} - \bar{S}_{R2})'\hat{\phi}_1$
Period 2 home and health inputs	$(\overline{H}_{U2} - \overline{H}_{R2})'\hat{\gamma}_1 + (\overline{h}_{U2} - \overline{h}_{R2})\hat{\varphi}_1$
Period 2 predetermined direct influences and period 2 omitted inputs	$(\bar{f}_U - \bar{f}_R)'\hat{\hat{\pi}} + (\bar{z}_U - \bar{z}_R)'\hat{\hat{\psi}} + \hat{\hat{\theta}}$

This decomposition strategy also allows for a simple test for omitted inputs. The intuition is that exogenous input determinants will contribute information to the estimation of a test score gap if skill inputs have not been fully accounted for. In particular, rejection of the null hypothesis $(\bar{z}_U - \bar{z}_R)'\psi + \theta = 0$ in the cumulative specification implies the presence of at least one omitted input. Rejection of the null $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} + \tilde{\theta} = 0$ in the value added specification implies the presence of at least one omitted period 2 input. Furthermore, if the null is rejected and we assume that variables in f_i have a positive effect on the demand of omitted inputs, then $(\bar{z}_U - \bar{z}_R)'\hat{\psi} + \hat{\theta}$ or $(\bar{z}_U - \bar{z}_R)'\hat{\hat{\psi}} + \hat{\hat{\theta}}$ (depending on the specification being used), can provide a lower bound for the contribution of the omitted inputs.[8]

At this point is worth discussing the main assumptions required for the above strategy to identify the contributions of early childhood and school influences to the cognitive gap under analysis. First, we are assuming that the production function of skill is linear. This will allow a direct comparison with the standard OB technique but comes at the cost of ignoring potential complementarities between inputs.

One type of complementarity potentially relevant for this analysis is what the literature refers to as “dynamic complementarity” (Cunha and Heckman, 2007). This means that cognitive skill attained during early childhood could increase the productivity of school inputs, so ignoring it could lead to an underestimation of the contribution of early childhood influences. One way of allowing for this type of complementarity is by adding interactions between school inputs and period 1 cognitive skill. We tested this and none of the interaction terms were statistically significant, both independently and jointly (see Table 4.1 in Appendix 4).[9]

Second, we are assuming that, after controlling for period 1 cognitive skill and input determinants contained in f_i, z_i and G_i, there are no remaining unobservables correlated with the inputs of interest.[10] Potential unobservables present in the error term of the hybrid value-

[8] Notice that the complete contribution of the omitted inputs is given by the contribution through $\bar{z}_U - \bar{z}_R$ plus the contribution through $\bar{f}_U - \bar{f}_R$. The contribution through $\bar{z}_U - \bar{z}_R$ can provide a lower bound for the contribution of the omitted inputs if the contribution through $\bar{f}_U - \bar{f}_R$ is positive.

[9] Future research focused on the contribution of other input categories such as contemporaneous home or health influences can evaluate interactions involving these inputs. It should be noticed that if an interaction is found to be significant, the decomposition strategy must include a special category to accommodate the contribution of concurrent differences in the inputs involved in the interaction.

[10] Notice that we could relax this assumption and still identify the contributions of interest. This assumption is required for the consistent estimation of production function parameters and this is not a necessary condition for the consistent estimation of the contribution of a

added specification (e_{i2}^{VA}, see equation (20)) are omitted period 2 inputs, the measurement error of lagged skill and the contemporaneous influence of innate ability. To rule out the presence of omitted period 2 inputs in the error term we are assuming that a linear combination of input determinants f_i, z_i and G_i is a sufficient representation of the demand function of omitted inputs.

We tested this by removing significant school inputs and evaluating whether their contribution was captured by the category that should be hosting omitted influences through the arguments of their demand functions. As predicted by the model, practically all of the contribution of the intentionally omitted school inputs ended up captured by the "period 2 predetermined direct influences and period 2 omitted inputs" category. This is shown and further discussed in Section 4.2. We also tested more flexible representations of the demand function of omitted inputs by allowing interactions between input determinants. Our main decomposition results proved robust to these specifications (see Figure 4.2 in Appendix 4).

Other potential sources of endogeneity that can be present in the error term of the value added specification are the measurement error of lagged skill and the contemporaneous influence of innate ability. In this regard, evidence reviewed and discussed in Singh (2015) indicates that these elements do not introduce a significant bias in the estimated effect of contemporaneous inputs. In fact, several studies show that value added models such as the one presented in (20) outperform other empirical strategies when recovering teacher effects from simulated data

category of inputs. In Appendix 3, we show that one can still obtain a consistent estimate of the contribution of a category of inputs despite some of them remain unobserved and produce a bias in the estimates of production function parameters. Fortin et al. (2011) claim that correlation between the error term and covariates can still allow one to obtain consistent estimates of the "unexplained" part of a BO decomposition as long as the dependence structure is the same in the two groups under analysis. The decomposition proposed here, however, requires separating the "explained" part of the gap into two different categories. As shown in Appendix 3, the assumption required for identification in this case is that the information contained in the observed inputs of a category suffices to predict the average urban/rural difference of the unobserved inputs that belong to the same category.

(Guarino et al., 2012), and provide the same results as experimental and quasi-experimental methods used to identify school or teacher effects (Deming et al. (2014) Kane et al. (2013), among others). Moreover, value added estimates given in Singh (2015) for the effect of private school enrolment on the achievement of rural children are similar to the results provided by an experimental exercise carried out in the same region of India (Muralidharan and Sundararaman, 2013).[11]

Our empirical strategy does not necessarily solve all endogeneity issues and there can still be room for concerns related to the endogeneity of school inputs. Because of this, to add reliability to our results, we will also perform a decomposition exercise using a value added specification after removing the correlation between lagged test scores and measurement error.[12] Because the potential biases affecting the estimate of the persistence parameter (ρ) caused by innate ability and measurement error operate in opposite directions, one can think of this estimation as an upper bound for the contribution of early childhood influences (only the possibility of a positive bias remains in our estimate of ρ) and a lower bound for the contribution of school inputs.

It is worth noticing that we are not using a restricted value added model. This means that we are not imposing the restriction $\rho = 1$ and, thus, we are not modeling the first difference of

[11] An empirical strategy that accounts for the presence of innate ability in the error term of the value added model could improve identification. A dynamic panel data approach could serve to differentiate out innate ability from this error term but is not applicable in this case. This is because we lack baseline (beginning of period 1) test scores and also because it will not allow us to recover parameter estimates of time invariant influences and, among these, those related to the school environment. With three rounds of test scores, Andrabi et al. (2011) proposed a dynamic panel data strategy to identify the private school premium in Pakistan and relied on school switching to recover this parameter. It is worth mentioning that they found estimates very similar to those provided by a value added model.

[12] This can be achieved by instrumenting lagged test scores (T_{i1}) with another measure of cognitive skill. The objective is to induce variation in T_{i1}uncorrelated with its own random measurement error (ε_{i1}). Notice this rests on the assumption of no correlation between the measurement errors affecting the two test scores.

skill measures between periods 2 and 1. If we impose this restriction we would be following a difference-in-differences approach, assuming that skill does not depreciate and not allowing for different trends prior to the first measurement of skill. As discussed in Andrabi et al. (2011), this would lead to an underestimation of the importance of contemporaneous influences if the true process has an autoregressive nature as expressed in (20).[13]

The third key assumption behind our decomposition strategy is that the geographical domain is not part of families' choices during the period under analysis. In fact, this strategy relies on classifying the indicator function for the groups of interest (urban/rural in our case) as part of the arguments of an input demand function and, thus, on assigning its contribution to the category hosting omitted inputs. This means we are assuming that families' choices regarding skill inputs are made taking their geographical domain as given. Consistent with this, more than 95% of the families in the sample remain in the same geographical domain between the three rounds of the survey considered for this analysis.

If one relaxes this assumption, it would be difficult to justify the inclusion of the urban/rural indicator in the empirical specifications ((19) and (20)) because it would not have a clear role as an input determinant. If the urban/rural indicator is not included, the sum of the individual contributions will not necessarily add-up to the difference in mean outcomes between the urban and rural domain (i.e. equations (21) and (22) will not necessarily hold). In addition, it would not be possible to directly relate our decomposition strategy to the BO technique and compare our results with those obtained using the rules employed in the literature. This is further explained below.

[13] The reader can also refer to Angrist and Pischke (2009) (Appendix 5.4) on the difference between assuming a standard fixed effect model (e.g. $T_{i2} = x_{i2}'\beta_1 + \alpha_i + \varepsilon_{i2}$) and an autoregressive process (e.g. $T_{i2} = \rho T_{i1} + x_{i2}'\beta_1 + \varepsilon_{i2}$), and the biases that will arise when estimating treatment effects (β_1) by first differencing when the true process is the latter. Notice that an autoregressive process is consistent with the cumulative nature of skill formation.

It is useful to relate our decomposition strategy to the BO decomposition so it can be compared to the strategies employed thus far in the literature. There are different ways to implement the BO decomposition. One can choose between a "threefold" or a "twofold" decomposition, and one needs to decide which will be the reference group used to measure the difference in coefficients that produces the "unexplained" part of the gap (Biewen, 2012; Elder et al., 2010; Jann, 2008). This "unexplained" part, in turn, can be interpreted as capturing a difference in returns to inputs or the presence of omitted inputs (McEwan and Trowbridge, 2007).

The decomposition strategy described here can be considered as a special case of "twofold" BO decomposition. In fact, inclusion of the group indicator G_i in the hybrid models described above ensures we are using the same coefficient estimates than those required to build the BO decomposition that uses the coefficients of a pooled regression as the reference coefficients. In addition, this can be regarded as a BO decomposition where the "unexplained" part of the gap is interpreted as capturing the contribution of omitted inputs.[14]

The distinctive feature of this decomposition strategy with respect to those employed thus far is that it is based on the results of a model that postulates a relation between cognitive skill and its inputs, and describes how families' choices determine these inputs. This prevents arbitrary choices of the reference group and arbitrary interpretations of the "unexplained" part

[14] Consider two groups of individuals (A and B) for whom a certain outcome (y_i) can be related to a set of predictors (x_i) in the following way: $y_{iA} = x'_{iA}\beta_A + \varepsilon_{iA}$ and $y_{iB} = x'_{iB}\beta_B + \varepsilon_{iB}$. Now consider the following pooled regression: $y_i = x'_i\beta_{Pool} + \delta D_i + \varepsilon_i$ where where D_i is a group indicator (it takes the value of 1 if the individual belongs to group A and the value of 0 if he belongs to group B). The inclusion of this group indicator ensures that $\bar{y}_A - \bar{y}_B = (\bar{x}_A - \bar{x}_B)'\,\hat{\beta}_{Pool} + \;\hat{\delta} = (\bar{x}_A - \bar{x}_B)'\,\hat{\beta}_{Pool} + \bar{x}'_A\big(\hat{\beta}_A - \hat{\beta}_{Pool}\big) + \bar{x}'_B\big(\hat{\beta}_{Pool} - \hat{\beta}_B\big)$. The term $\hat{\delta} = \bar{x}'_A\big(\hat{\beta}_A - \hat{\beta}_{Pool}\big) + \bar{x}'_B\big(\hat{\beta}_{Pool} - \hat{\beta}_B\big)$ corresponds to the "unexplained" part of the gap in the BO decomposition that uses the coefficients of the pooled regression as reference coefficients (Elder et al., 2010). Notice, therefore, that we are not imposing the restriction of equal coefficients for both groups. What we are restricting is the interpretation of this coefficient difference (or "unexplained" part) to mean the presence of omitted inputs.

of the gap. Notice that, as opposed to a more standard BO decomposition, in this strategy the group indicator is not the only variable accounting for omitted inputs. Predetermined direct influences and other input determinants are also included in the category hosting omitted inputs because, according to the model, they all have a role as arguments in their demand functions.

To illustrate the risk of bias if one follows the decomposition strategy employed in the literature, we will also estimate the contributions of early childhood influences and school inputs using the value added specification and a standard decomposition rule. In particular, we will assign all household, family and child characteristics to the category hosting period 2 home and health inputs. Following a standard BO decomposition, the coefficient of the group indicator will capture the "unexplained" part of the gap (see Table 3). In the following section we present and discuss the results obtained using this decomposition rule.

Table 3
Decomposition categories and variables included under the standard decomposition rule

Value added specification	
Early childhood influences	$(\bar{\bar{T}}_{U1} - \bar{\bar{T}}_{R1})\hat{\rho}$
School inputs	$(\bar{\bar{S}}_{U2} - \bar{\bar{S}}_{R2})'\hat{\phi}_1$
Period 2 home and health inputs	$(\bar{H}_{U2} - \bar{H}_{R2})'\hat{\gamma}_1 + \left(\bar{f}_U - \bar{f}_R\right)'\hat{\hat{\pi}} + (\bar{z}_U - \bar{z}_R)'\hat{\hat{\psi}} + \left(\bar{h}_{U2} - \bar{h}_{R2}\right)\hat{\varphi}_1$
Unexplained	$\hat{\hat{\theta}}$

4. Data and decomposition results

4.1. Data sources and variables

This analysis will employ the information contained in the first three rounds of the child and household surveys, as well as the school survey, focusing on the Younger Cohort of the Young Lives Study in Peru.[15] The basic structure of this data is summarized in Table 4.

The estimations will be based on two different samples. The first considers all children that have cognitive test scores for rounds 2 and 3, and attend a school included in the school survey (487 children in 124 schools).[16] The second sample considers all children that have cognitive test scores for rounds 2 and 3 (1,561 children).

Table 4
Structure and sample sizes of the relevant Young Lives databases

	Child and household survey			*School Survey 2011*
	Round 1 2002	*Round 2 2006*	*Round 3 2009*	
Younger cohort's age (years)	1 (0.5-1.5)	5 (4.5-5.5)	8 (7.5-8.5)	10 (9.5-10.5)
Sample size (children)	2,052	1,963	1,943	572 (132 schools)
Educational attainment	--	Preschool	Grade 2	Grade 5

Source: Young Lives Study (Peru).

Following the analytical framework described in Section 2, period 1 variables will correspond to influences captured in rounds 1 and 2 (at ages 1 and 5) and period 2 variables will correspond to influences captured in round 3 (at age 8). Influences captured in the school

[15] Young Lives is an international study of childhood poverty, following 12,000 children in 4 countries (Ethiopia, India, Peru and Vietnam) over 15 years.

[16] The risk of selection bias due to this second condition is very small. Primary school attendance in Peru is close to 100% (only 0.7% of Young Lives younger cohort children were not attending school in Round 3) and schools participating in the school survey were randomly selected (Guerrero et al., 2012). Also, the second requirement for inclusion will be relaxed in some of the specifications to explore whether the results are affected by restricting the sample to those children whose school was included in the school survey.

survey (collected two years after round 3) will be assumed to be the same as those present in period 2. We are also assuming that the child has remained in the same school since her enrolment in Grade 1 (at age 6) until the school survey was conducted (at age 10).[17]

Table 5 presents all the variables considered for the analysis. The measures of cognitive achievement employed are the standardized test scores obtained in the Peabody Picture Vocabulary Test (PPVT). This is a widely used test of receptive vocabulary that has a strong positive correlation with several measures of intelligence (Cueto and Leon, 2012).

It is possible to use the information collected in round 2 to estimate expenditure flows invested in the child and to identify her access to preschool education. The survey, however, is not very informative of the quality of care and the home environment during infancy (round 1). Mothers' access to antenatal care (where advice on parenting practices is usually provided) and the way mothers' responded to the child crying were used to approximate the quality of care provided during early childhood.

Period 2 (round 3) information regarding educational home inputs is richer. In addition to the expenditure flows invested in the child, it is possible to account for the child's access to learning materials and resources such as books and computers, parental engagement in educational activities (i.e. providing help with homework), and the amount of time the child devoted to studying at home.

The variable chosen to reflect health inputs is an indicator of whether the child was stunted or not. This provides a fairly objective summary indicator of the child's health status. In addition, the causal relation between this measure of nutritional status and cognitive skill has already been documented (Outes-Leon et al., 2011).

[17] According to administrative data collected from the schools included in the survey, school switching is not significant. On average, only 2% of the students enrolled in primary education changed school between 2009 and 2010.

Following the classifications proposed in Guerrero et al. (2012), the information contained in the school survey was grouped into six categories: (i) school size, organization and timetable; (ii) infrastructure; (iii) climate[18]; (iv) learning activities and materials; (v) teacher characteristics; and (vi) school responsiveness[19]. School variables presented in Table 6 are the ones which resulted after applying a three-step procedure to reduce the amount of noise and narrow down the most significant predictors of cognitive skill among the school inputs available.[20]

Predetermined direct influences refer to variables that are outside the current choice set of families and that have a direct effect on skill. These include caregiver´s educational attainment and age, and child characteristics such as sex, age and mother tongue. These child characteristics can accommodate biological differences between children as well as potential advantages due to language when completing the PPVT (the test was administered in Spanish).

Finally, input determinants include variables reflecting family resources, parental preferences, child and sibling characteristics that can affect parental investments, and an urban/rural indicator. As already explained, this last variable is intended to capture

[18] Variables in this category include teachers' perception of the quality of relations among students and between students and teachers, and of the problems and difficulties encountered during the school year.

[19] Variables within this category indicate whether or not the school provides support for students lagging behind or at risk of dropping out.

[20] The three-step procedure was as follows: (i) pairwise correlations between candidate variables within each category were evaluated, variables with correlation coefficients below 0.6 were chosen and those with a correlation above 0.6 with two or more others were discarded; (ii) a regression of PPVT scores on the variables chosen after (i) was run for each category, and variables with a significant partial correlation were chosen; and (iii) a regression of PPVT scores on the variables chosen after (ii) was run, and those with a significant partial correlation were chosen. The results reported in the next section are robust to using the first two principal components of the six school quality dimensions (see Figure 4.3 in Appendix 4).

differences in the general health environment and in the availability of educational goods and services in the two geographical domains.[21]

Table 5 presents descriptive statistics as well as urban/rural differences for all the variables described above. Significant positive differences between urban and rural children are present in most of the direct influences and input determinants considered. This corroborates what has already been established by several studies about the Peruvian basic education system: there are high levels of enrolment but school quality remains very heterogeneous and unequally distributed between children of different socioeconomic backgrounds (Beltran and Seinfeld, 2012; Cueto et al., 2014) leading to a highly segregated system.

[21] Consistent with the assumption that input determinants are not affected by families' choices during the period under analysis, more than 95% of the families in the sample remain in the same geographical domain between rounds 1 and 3.

Table 5
Description of the variables used in the empirical specifications

Variable type	Variable used in empirical specifications	Round	Mean	SD	Urban	Rural	Diff.
Period 1 measured cognitive skill (T_{i1})	Standardized raw PPVT score	2	1.780	0.951	2.095	1.028	1.067*** (0.14)
Period 2 measured cognitive skill (T_{i2})	Standardized raw PPVT score[(a)]	3	0.024	0.968	0.355	-0.766	1.121*** (0.13)
Early childhood educational inputs (H_{i1})	Real expenditure in child (learning materials and entertainment; x1,000 soles; 2006 prices in urban Lima)	2	0.274	0.364	0.342	0.112	0.23*** (0.049)
	Mother had antenatal visits during pregnancy (yes = 1)	1	0.828	0.378	0.848	0.778	0.071* (0.038)
	Maternal response to child cry was affectionate (yes = 1)[(b)]	1	0.230	0.421	0.286	0.097	0.188*** (0.05)
	Child attended formal preschool (yes = 1)	2	0.766	0.424	0.892	0.465	0.427*** (0.055)
Period 2 educational home inputs (H_{i2})	Real expenditure in child (learning materials and entertainment; x1,000 soles; 2006 prices in urban Lima)	3	0.432	0.572	0.517	0.230	0.287*** (0.063)
	Household has books and child is encouraged to read (yes = 1)	3	0.450	0.498	0.478	0.382	0.096 (0.06)
	Household has a computer (yes = 1)	3	0.140	0.347	0.195	0.007	0.188*** (0.039)
	Child receives help from parents when doing homework (yes = 1)	3	0.665	0.472	0.758	0.444	0.314*** (0.029)
	Hours in a typical day the child spends playing	3	4.346	1.517	4.488	4.005	0.483** (0.218)
	Hours in a typical day the child spends sleeping	3	9.931	0.978	9.988	9.796	0.192 (0.114)
	Hours in a typical day the child spends studying	3	1.945	0.834	2.120	1.526	0.594*** (0.078)
Period 1 health input (h_{i1})	Child is stunted (yes = 1)[(c)]	2	0.316	0.465	0.207	0.576	-0.369*** (0.034)
Period 2 health input (h_{i2})	Child is stunted (yes = 1)	3	0.189	0.392	0.120	0.354	-0.235*** (0.041)

Variable type	Variable used in empirical specifications	Round	Mean	SD	Urban	Rural	Diff.
School inputs (S_{i2})	Years of schooling (basic education)	3	2.374	0.544	2.429	2.243	0.186** (0.085)
	Hours in a typical day the child spends at school[(d)]	3	6.171	0.720	6.131	6.269	-0.138 (0.108)
	CLIM: absence of problems is class (score 12-48)[(e)]	School survey	32.736	6.567	33.760	30.298	3.462** (1.317)
	INF: school has basic services (yes = 1)[(f)]	School survey	0.556	0.497	0.761	0.069	0.691*** (0.087)
	ACT: average curricular coverage in maths and language (average % of topics covered in depth) [(e)]	School survey	0.531	0.153	0.564	0.452	0.111*** (0.034)
	ORG: teacher absenteeism (%)[(g)]	School survey	0.025	0.111	0.012	0.057	-0.045 (0.031)
	ORG: school has a psychologist (yes = 1)	School survey	0.179	0.383	0.248	0.014	0.234* (0.109)
	ORG: school is "multigrade" (yes = 1)[(h)]	School survey	0.187	0.390	0.073	0.458	-0.385*** (0.084)
	TEA: more than 50% of teachers graduated from a university (yes = 1)[(e)]	School survey	0.456	0.499	0.551	0.229	0.322*** (0.091)
Predetermined direct influences (f_i)	Child's caregiver has higher education (yes = 1)	3	0.179	0.383	0.245	0.021	0.224*** (0.037)
	Caregiver's age	3	34.569	6.843	34.172	35.514	-1.342 (0.804)
	Child is male (yes = 1)	3	0.478	0.500	0.490	0.451	0.038 (0.048)
	Child's mother tongue is Spanish (yes = 1)	3	0.893	0.309	0.985	0.674	0.312** (0.104)
	Child's age in months	3	96.510	3.708	96.500	96.537	-0.037 (0.507)
Exogenous input determinants (z_i)	Child lives in urban area (yes = 1)	3	0.704	0.457	1.000	0.000	1.000
	Average household total income (x10,000 soles; 2006 prices in urban Lima)	2 and 3	1.512	1.116	1.711	1.037	0.674*** (0.111)
	Average household size	1, 2 and 3	5.538	1.849	5.270	6.176	-0.906** (0.306)

Variable type	Variable used in empirical specifications	Round	Mean	SD	Urban	Rural	Diff.
	Proportion of male siblings	1, 2 and 3	0.495	0.333	0.490	0.506	-0.016 (0.026)
	Child birth order	1, 2 and 3	2.475	1.584	2.194	3.144	-0.949*** (0.198)
	Caregiver aspiration for child's educational attainment is university education (yes=1)	2 and 3	0.655	0.476	0.743	0.444	0.299*** (0.065)

(a) Round 3 and round 2 raw PPVT scores were standardized using the round 2 mean and standard deviation.
(b) Mother cuddled or soothed child when he/she cried.
(c) A child is considered stunted if she exhibits a height for age z score below -2.
(d) The effects of children's time use categories are measured with respect to time spent working (the omitted time use category).
(e) Reported by mathematics and language teachers in charge of classes attended by Young Lives children. Problems are related to student absenteeism, lack of motivation, discipline and peer relations.
(f) Basic services comprise water (from a public network or pipe), sanitation (public network connection or a treated cesspool), electricity and telephone connection.
(g) Measured by observation, in maths and language classes attended by Young Lives children.
(h) "Multigrade" means that children from different grades receive classes at the same time, in the same room, and by the same teacher.
The number of observations is 487 for all variables.
Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

4.2. Decomposition results and discussion

The empirical goal is to estimate the contribution of early childhood and school influences to the cognitive gap observed, at age 8, between urban and rural children in Peru. For this, we will rely on the results of the decomposition strategy described in Table 2 and all the information available from the Young Lives Study. We will call this the “full information” decomposition. We will privilege the value added specification because the results documented in the literature show that this specification can provide reliable estimates of the effect of contemporaneous influences on skill (Singh, 2015). The results of the cumulative model will also be considered because they will serve to illustrate the role of exogenous input determinants as variables that control for omitted inputs.

We will also empirically illustrate the main issues discussed above regarding the decomposition strategies employed in the literature. Namely: (i) that assigning all available household, family and child characteristics into a single category will likely lead to an overstatement of the importance of influences originated at home *vis-à-vis* that of school inputs, especially when one lacks rich information on school characteristics; and (ii) that the decomposition strategy proposed here is less prone to biases than those employed so far in the literature.

For this, we will perform two additional decompositions after excluding the school inputs contained in the school survey. The first will be based on the components of the new decomposition strategy and will serve to determine whether predetermined direct influences and exogenous input determinants pick-up the contribution of the omitted school inputs as predicted by the model described in Section 2. The second will be based on the components of the standard decomposition rule and will be compared against the “full information decomposition" to verify if this rule introduces a positive bias in the estimated contribution of family and household influences.

Table 6 presents the estimated contributions of each category: (i) including all inputs from the school survey (the "full information" decomposition; see Panel A); (ii) excluding the inputs from the school survey[22] and using the same sample of children as in (i) (see panel B); and (iii) excluding the inputs from the school survey and using the complete sample of children that register a PPVT score in rounds 2 and 3 (see panel C). Table 7 presents the results using the standard decomposition rule, excluding the inputs from the school survey, and using the complete sample of children.

Figure 3 shows the same point estimates as Table 6 accompanied by 95% confidence intervals. It also presents the statistic and corresponding p-value of the test of omitted inputs described in the previous section. Recall that this statistic provides an estimate of the contribution of exogenous input determinants to the gap under analysis. Appendix 2 presents the coefficient estimates of the variables involved in all the specifications.[23]

The first set of results reveals that school inputs have a significant contribution of around 35% to the cognitive skill gap observed, at age 8, between urban and rural children in Peru. To account for this contribution we are reporting the estimate provided by the value added specification. The cumulative hybrid model can control for omitted inputs but retains the full cumulate effect of unobserved innate ability. This model produces a larger contribution for school inputs, which is likely affected by a positive bias.

[22] Ignoring the information contained in the school survey implies that the only school inputs considered are years of schooling and time spent at school.

[23] The value added models were estimated including the period 1 inputs available. This does not alter the interpretation of its coefficients and is a less restrictive specification as it relaxes the assumption requiring that the effects of period 1 inputs decay at a rate ρ. The contributions of these period 1 inputs were assigned to the early childhood environment.

Results reported in panel A of Figure 3 show that exogenous input determinants make a significant contribution (22%; $p < 0.05$) in the cumulative specification. This indicates the presence of omitted inputs that are being controlled for through their demand equations. In the value added model, input determinants no longer make a significant contribution to the gap. In this model, we are controlling for all period 1 inputs through lagged test scores. Therefore, failure to reject the null $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$ in the value added model ($p = 0.27$) indicates the absence of period 2 omitted inputs.

If the contribution of omitted inputs is captured by their demand equations when exogenous input determinants are included in the regression, the omission of significant school inputs should increase the contribution of the "predetermined direct influences and omitted inputs" category. Results presented in Panel B of Table 6 and Figure 3 are consistent with this. In particular, the contribution of the category hosting omitted inputs in the cumulative specification grows twice as large when the school survey information is omitted (from 33% in the "full information" decomposition up to 65% in the decomposition that excludes school inputs).

There is also a significant increase in the contribution of the category hosting omitted inputs in the value added specification (from 21% in the "full information" decomposition up to 42% after ignoring the school inputs provided by the school survey). Importantly, the contribution of exogenous input determinants is now significant in the value added model (30%; $p < 0.00$) which implies we can reject the null $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} + \tilde{\theta} = 0$ (see Panel B in Figure 3). This result differs from the one obtained with the complete set of data and confirms that there are now omitted period 2 influences. This is consistent with the fact that we are intentionally omitting school inputs.

It is also worth noticing that there is only a small increase in the estimated contribution of the categories hosting early childhood and home inputs after excluding school influences. This increase remains well within standard errors in both specifications (compare panels A and B in Figure 3), which means that the demand equations are concentrating most of the contribution of the omitted school inputs.

Results presented in Panel C of Table 6 and Figure 3 reveal that the results just discussed are robust to considering the entire sample of children with complete PPVT scores and not just those attending schools included in the school survey. This should mitigate concerns regarding potential selection bias in the sample used for the preceding analysis. The use of a larger sample also adds precision to the results discussed in the previous paragraphs.[24]

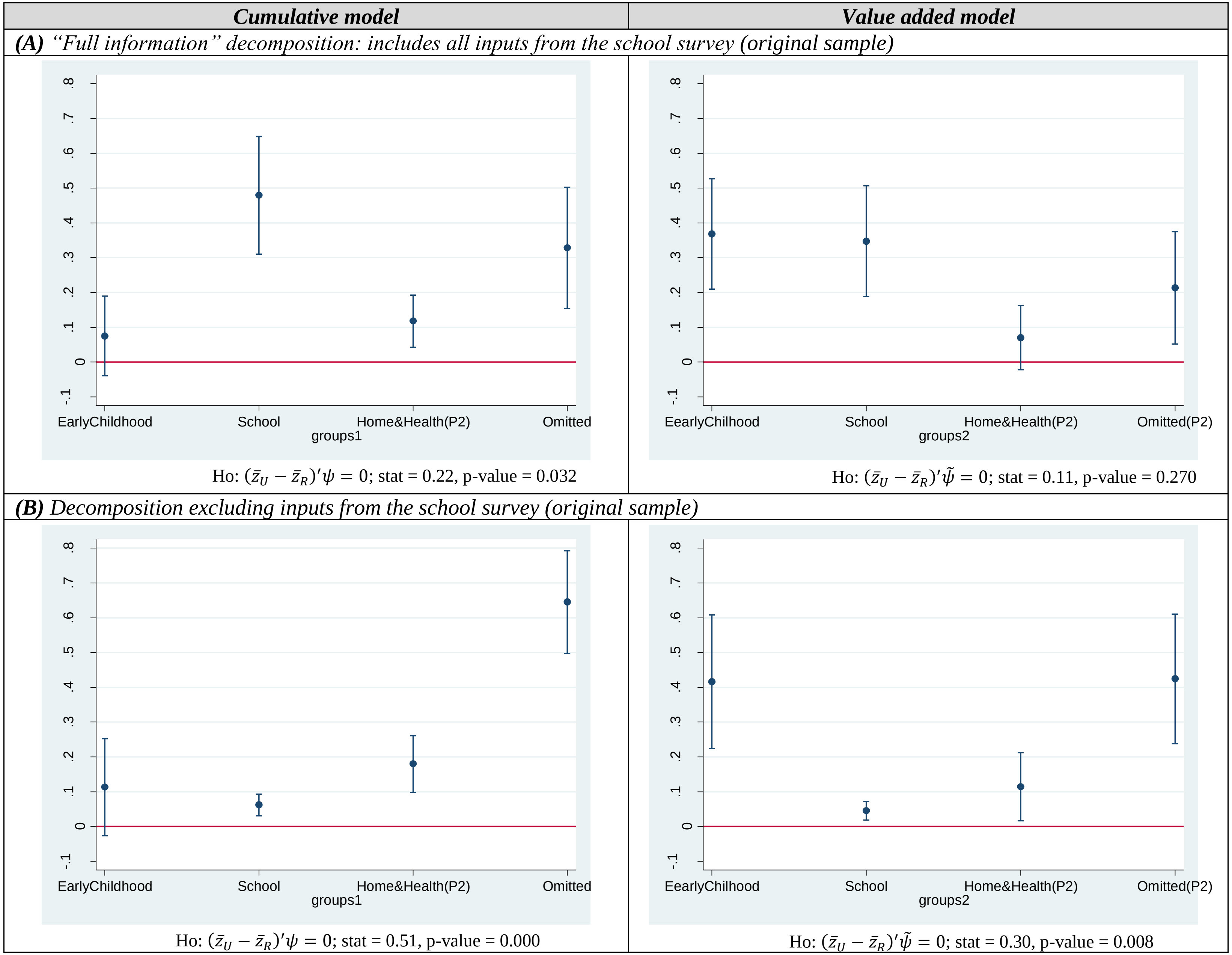
[24] We also attempted a decomposition exercise using the scores of a Mathematics test applied in Round 3 (see Cueto and Leon (2012) for details about this test). Unfortunately, the same test was not applied in round 2 and, therefore, we could only estimate a cumulative specification (see Table 4.4 in Appendix 4). The results obtained for the cumulative model are robust to the use of the Mathematics test scores, albeit less precise. Similar to the results obtained with the PPVT scores, the estimated contribution of school inputs is around 45%. As already discussed, this is likely affected by a positive bias because the cumulative specification retains the entire effect of innate ability in the error term. Importantly, after the school inputs are intentionally omitted, the category hosting omitted inputs also captures the majority of their contribution and this result is robust to the use of the complete sample of children (see Table 4.4 in Appendix 4).

Table 6
Normalized contributions to the urban/rural gap in cognitive skill at age 8

Cumulative specification		***Value added specification***	
(A) *“Full information” decomposition: includes all inputs from the school survey*			
Early childhood educational and health inputs	0.075 (0.058)	Early childhood influences	0.368*** (0.081)
School inputs	0.479*** (0.086)	School inputs	0.348*** (0.081)
Period 2 home and health inputs	0.118*** (0.038)	Period 2 home and health inputs	0.071 (0.047)
Predetermined direct influences and omitted inputs	0.328*** (0.089)	Period 2 predetermined direct influences and period 2 omitted inputs	0.213** (0.082)
(B) *Decomposition excluding inputs from the school survey (original sample)*			
Early childhood educational and health inputs	0.113 (0.071)	Early childhood influences	0.416*** (0.098)
School inputs	0.062*** (0.016)	School inputs	0.045*** (0.014)
Period 2 home and health inputs	0.180*** (0.042)	Period 2 home and health inputs	0.114** (0.050)
Predetermined direct influences and omitted inputs	0.645*** (0.075)	Period 2 predetermined direct influences and period 2 omitted inputs	0.424*** (0.095)
(C) *Decomposition excluding inputs from the school survey (complete sample)*			
Early childhood educational and health inputs	0.096** (0.025)	Early childhood influences	0.419*** (0.032)
School inputs	0.051*** (0.009)	School inputs	0.027*** (0.007)
Period 2 home and health inputs	0.165*** (0.030)	Period 2 home and health inputs	0.107*** (0.026)
Predetermined direct influences and omitted inputs	0.689*** (0.057)	Period 2 predetermined direct influences and period 2 omitted inputs	0.447*** (0.055)

Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Figure 3
Normalized contributions to the urban/rural gap in cognitive skill at age 8 (point estimates and 95% confidence intervals)



Cumulative model | ***Value added model***

***(A)** “Full information” decomposition: includes all inputs from the school survey (original sample)*

Cumulative model: Ho: $(\bar{z}_U - \bar{z}_R)'\psi = 0$; stat = 0.22, p-value = 0.032

Value added model: Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.11, p-value = 0.270

***(B)** Decomposition excluding inputs from the school survey (original sample)*

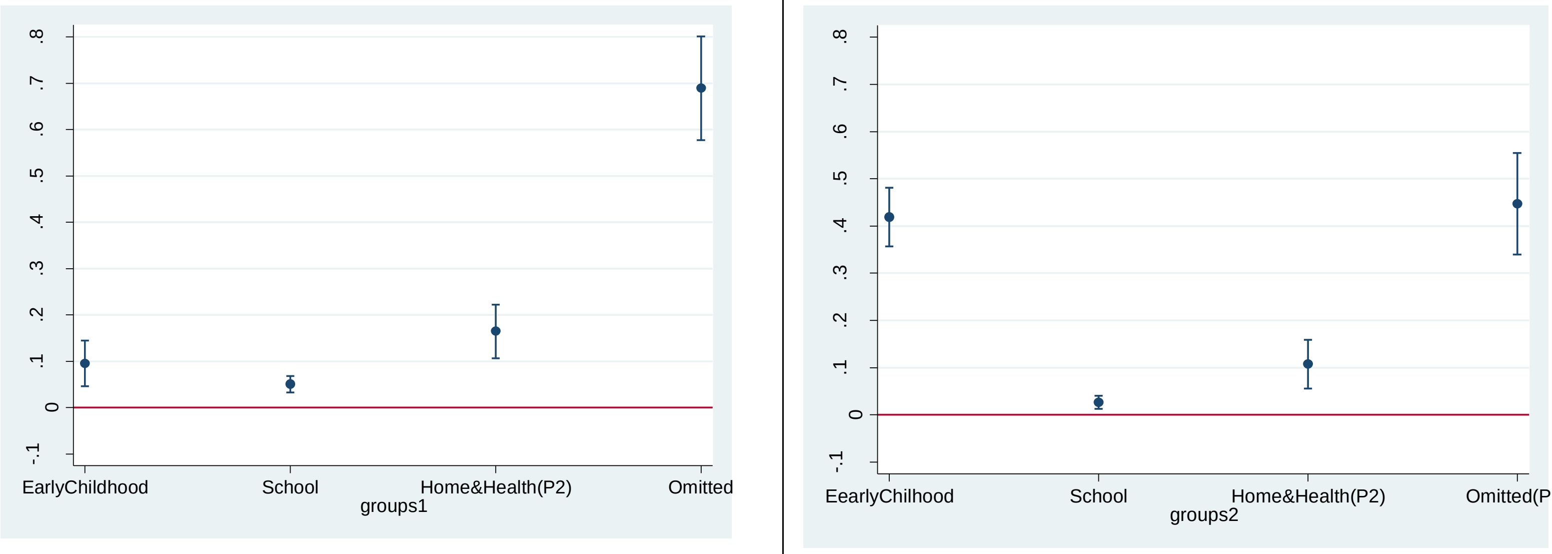
Cumulative model: Ho: $(\bar{z}_U - \bar{z}_R)'\psi = 0$; stat = 0.51, p-value = 0.000

Value added model: Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.30, p-value = 0.008

Figure 3 Normalized contributions to the urban/rural gap in cognitive skill at age 8 (point estimates and 95% confidence intervals)

Cumulative model	***Value added model***

***(C)** Decomposition excluding inputs from the school survey (complete sample)*



Ho: $(\bar{z}_U - \bar{z}_R)'\psi = 0$; stat = 0.56, p-value = 0.000

Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.32, p-value = 0.000

Notes:

“EarlyChildhood” refers to early childhood educational and health inputs for the cumulative model and refers to early childhood influences for the value added model.

“School” refers to school inputs.

“Home&Health(P2)” refers to period 2 home and health inputs.

“Omitted” refers to predetermined direct influences and omitted inputs.

“Omitted(P2)” refers to period 2 predetermined direct influences and period 2 omitted inputs.

Recall that the assumptions required to group all family, child and household characteristics in a single category are that there are no omitted inputs from the school environment or that school inputs are not affected by families' choices. The results discussed above show that none of these assumptions hold after we exclude the school inputs from the school survey. Therefore, if one follows the rule employed in the literature, it will likely generate a bias and overstate the contribution of the category hosting all family, child and household characteristics.

The results obtained after following this rule are shown Table 7. All family, household and child characteristics are assigned to the category hosting period 2 home and health inputs. As in a standard BO decomposition, the contribution of the urban/rural indicator is assigned to a category accounting for the "unexplained" part of the gap. Notice that the category hosting period 2 home and health inputs presents a contribution 26 percentage points larger than in the "full information" decomposition. We know at least part of this additional contribution is a bias because at least part of it belongs to the school environment through the school inputs we are intentionally omitting.

The strategy proposed here prevents one from incurring in the bias described above. Its results are shown in Panel C of Table 6 and Figure 3, under the value added specification. With these results one can conclude that early childhood influences account for around 40% of the gap in cognitive development and that there are omitted period 2 influences that account for, at least, 32% of the gap (if one uses the statistic testing for omitted inputs as a lower bound for their contribution). Unlike the conclusions derived from applying the standard decomposition rule, these conclusions are consistent with the results provided by the "full information" decomposition.

Table 7
Normalized contributions to the urban/rural gap in cognitive skill at age 8: standard decomposition rule and complete sample

Early childhood influences	0.419*** (0.032)
School inputs	0.027*** (0.007)
Period 2 home and health inputs	0.327*** (0.012)
Unexplained	0.228*** (0.077)

Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Finally, in Appendix 4 we also present decomposition results from a hybrid value added model after instrumenting lagged PPVT test scores with the results of the Cognitive Developmental Assessment (CDA) test, also administered to the younger cohort during the second round (see Cueto and Leon (2012) for details). The objective is to remove the potential attenuation bias due to measurement error from our estimates of the persistence parameter, and provide an upper bound for the contribution of early childhood influences and a lower bound for the contribution of school inputs. As expected, the instrumental variable estimate of the persistence parameter is larger than its OLS counterpart (compare columns (2) and (3) in Appendix 2). Accordingly, the contribution of past influences is now situated around 50% while school inputs contribute 28% of the gap (see Table 4.5 in Appendix 4).

5. Concluding remarks

Children of dissimilar socioeconomic backgrounds demonstrate significant differences in cognitive outcomes across the developing world. Cognitive skill formation is a cumulative process and, therefore, influences that have taken place early in the life of these children but also later, at school, can both play in role in shaping these gaps.

This paper sought to contribute to the literature by using a new decomposition strategy to measure the contribution of early childhood and school influences to the cognitive gap observed between urban and rural eight-year-olds in Peru.[25] We argued and empirically illustrated how this decomposition strategy is less prone to biases than those employed so far in the literature.

We found that school influences occurring between ages 6 and 8, account for a significant share of urban/rural cognitive gap (around 35% and no less than 28%). The share attributable to early childhood influences is important but no larger than 50%.

This result has an important implication for the interpretation of evidence regarding the evolution of cognitive achievement gaps. Cognitive gaps usually emerge before children start school and remain fairly unchanged throughout their school years. This has been put forward by some authors as evidence that schools contribute little to these gaps and to cognitive development in general (Cunha et al., 2006; Heckman, 2006).

Judging the contribution of schools by comparing the size of achievement gaps at school-age against the size of gaps present during early childhood, however, relies on the assumption that skill exhibits perfect persistence (i.e. that there is no depreciation or that $\rho = 1$). The results presented here do not support this assumption. In fact, our estimate of the persistence parameter is between 0.34 and 0.52. This means that only a fraction (between 70 and 80%) of the gap is carried forward to the next period.[26] It also

[25] The objective of this analysis was to measure the contribution of early childhood and school inputs to the urban/rural cognitive gap. Our main decomposition results, however, also considered the contribution of contemporaneous home and health inputs and of predetermined direct influences and omitted inputs. According to our analytical framework, these two additional categories are necessary for a complete account of the cognitive skill gap.

[26] Considering that 3 years have passed between the two measurements of skill and that $0.7^3 = 0.34$. Notice that we are assuming that the rate of depreciation is the same for urban and rural children. In fact, we are assuming that all production function

means that the difference in cognitive skill at age 8 is similar to that observed at age 5 not because schools do not matter but because inequalities in school environments are sustaining a gap that would otherwise be smaller.

The characteristics of rural schools and teachers have a direct connection with policy action because nearly all the supply of educational services in rural Peru is public. Therefore, our results have a clear policy implication: efforts devoted to the equalization of the characteristics of rural schools and teachers with those existing in urban areas can produce a significant reduction in the cognitive skill gap between urban and rural children by the time they reach grade 3.

Our decomposition strategy was devised to identify the overall contribution of early childhood and school influences and not the contribution of a particular input. Therefore, our results cannot directly inform policy about a specific school or teacher characteristic to prioritize in the effort of equalizing school services between rural and urban areas. Inspection of individual coefficients in the three main specifications considered, however, suggests some school characteristics which warrant further research in terms of their potential influence in compounding the educational disadvantage of rural students. In particular, the use of multi-grade classrooms, teacher absenteeism, and the incidence of teacher-reported problems in class (related to student absenteeism, lack of motivation, discipline and peer relations) stand out as good predictors of cognitive test scores (see columns (1)-(3) in Appendix 2).

parameters are the same between the urban and rural domains. One way of interpreting this is that we are assuming that there are no biological differences between urban and rural children that affect their learning processes. A similar depreciation rate implies that more skilled (urban) children experience greater depreciation losses in absolute terms than less skilled (rural) children. If we allow for interactions between inputs and past skill, the productivity of past skill will no longer be limited between zero and one, and could vary across children. As already discussed, this possibility was tested and the interactions were not found significant.

Regarding the use of multi-grade classrooms, available causal evidence suggests there is nothing inherently prejudicial for learning about this type of environments. In fact, results for developed countries are mixed showing both positive (Thomas, 2012) and negative (Sims, 2008) effects, while more recent evidence shows that the overall effect of grade mixing can be positive, negative, or zero depending on the peer composition of the class (Leuven and Ronning, 2014). The potentially harmful characteristic of multi-grade classrooms in developing countries (including rural Peru) is that these settings are usually the consequence of scarce school resources so teachers lack the materials and training to manage this type of environments.

References

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

Angrist, J., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*: Princeton University Press.

Arteaga, I., & Glewwe, P. (2014). *Achievement Gap between Indigenous and Non-Indigenous Children in Peru: An Analysis of Young Lives Survey Data*. Young Lives Working Paper 130.

Beltran, A., & Seinfeld, J. (2012). *La Trampa Educativa en el Peru: Cuando la Educaion Llega a Muchos pero sirve a Pocos*. Peru: Universidad del Pacifico.

Biewen, M. (2012). *Additive Decompositions with Interaction Effects*. IZA DP No. 6730. Institute for the Study of Labor.

Blinder, A. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human Resources, 8*(4), 436-455.

Cueto, S., Guerrero, G., Leon, J., Zapata, M., & Freire, S. (2014). The relationship between socioeconomic status at age one, opportunities to learn and achievement in mathematics in fourth grade in Peru. *Oxford Review of Education, 40*, 50-72.

Cueto, S., & Leon, J. (2012). *Psychometric Characteristics of Cognitive Development and Achievement Instruments in Round 3 of Young Lives*. Young Lives Technical Note 25.

Cunha, F., & Heckman, J. (2007). The Technology of Skill Formation. *American Economic Review, 97*(2), 31-47.

Cunha, F., Heckman, J., Lochner, L., & Masterov, D. (2006). Interpreting the Evidence on Life Cycle Skill Formation. In E. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (pp. 697–812). Amsterdam: North-Holland.

Deming, D., Hastings, J., Kane, T., & Staiger, D. (2014). School Choice, School Quality, and Postsecondary Attainment. *American Economic Review, 104*(3), 991–1013.

Elder, T., Goddeeris, J., & Haider, S. (2010). Unexplained gaps and Oaxaca–Blinder decompositions. *Labour Economics, 17*, 284-290.

Fortin, N., Lemieux, T., & Firpo, S. (2011). Decomposition Methods in Economics. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. 4A): Elsevier.

Glewwe, P., & Miguel, E. (2008). The Impact of Child Health and Nutrition on Education in Less Developed Countries. In T. P. Schultz & J. A. Strauss (Eds.), *Handbook of Development Economics* (Vol. 4): Elsevier.

Grantham-McGregor, S., Cheung, Y. B., Cueto, S., Glewwe, P., Richter, L., Strupp, B., & Group, I. C. D. S. (2007). Developmental potential in the first 5 years for children in developing countries. *The Lancet, 369*(9555), 60-70.

Guarino, C., Reckase, M. D., & Wooldridge, J. M. (2012). *Can value-added measures of teacher performance be trusted?* IZA Discussion Papers 6602.

Guerrero, G., Leon, J., Rosales, E., Zapata, M., Freire, S., Saldarriaga, V., & Cueto, S. (2012). *Young Lives School Survey in Peru: Design and Initial Findings*.

Heckman, J. (2006). Skill Formation and the Economics of Investing in Disadvantaged Children. *Science, 312*, 1900-1902.

Heckman, J. (2007). The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Science, 104*(33), 13250-13255.

Hernandez-Zavala, M., Patrinos, H., Sakellariou, C., & Shapiro, J. (2006). *Quality of Schooling and Quality of Schools for Indigenous Students in Guatemala, Mexico and Peru*. WPS3982. World Bank.

Jann, B. (2008). The Blinder–Oaxaca decomposition for linear regression models. *The Stata Journal, 8*, 453-479.

Kane, T., McCaffrey, D., Miller, T., & Staiger, D. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Research Paper. MET Project. Technical report. Bill & Melinda Gates Foundation.

Leuven, E., & Ronning, M. (2014). Classroom Grade Composition and Pupil Achievement. *The Economic Journal, 26*(1), 1164-1192.

McEwan, P. (2004). The Indigenous Test Score Gap in Bolivia and Chile. *Economic Development and Cultural Change, 53*, 157-190.

McEwan, P., & Marshall, J. (2004). Why Does Academic Achievement Vary Across Countries? Evidence from Cuba and Mexico. *Education Economics, 12*(3), 205-217.

McEwan, P., & Trowbridge, M. (2007). The achievement of indigenous students in Guatemalan primary schools. *International Journal of Educational Development, 27*, 61-76.

Muralidharan, K., & Sundararaman, V. (2013). *The aggregate effect of school choice evidence from a two-stage experiment in India*. Working Paper 19441. National Bureau of Economic Research.

Oaxaca, R. (1973). Male–female wage differentials in urban labor markets. *International Economic Review, 14*(3), 693-709.

Outes-Leon, I., Porter, C., & Sánchez, A. (2011). *Early Nutrition and Cognition in Peru: A Within-Sibling Investigation*. IDB Working Paper Series No. IDB-WP-241. Inter-American Development Bank.

Paxson, C., & Schady, N. (2007). Cognitive Development among Young Children in Ecuador The Roles of Wealth, Health and Parenting. *The Journal of Human Resources, XLII*(1), 49-84.

Ramos, R., Duque, J. C., & Nieto, S. (2012). *Decomposing the Rural-Urban Differential in Student Achievement in Colombia Using PISA Microdata*. IZA DP No. 6515. Institute for the Study of Labor.

Rosenzweig, M., & Schultz, P. (1983). Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs, and Their Effects on Birth Weight. *Journal of Political Economy, 91*(5), 723-746.

Schady, N., Behrman, J., Araujo, M. C., Azuero, R., Bernal, R., Bravo, D., . . . Vakis, R. (2014). *Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries*. IDB-WP-482. Inter-American Development Bank.

Sims, D. (2008). A Strategic Response to Class Size Reduction: Combination Classes and Student Achievement in California. *Journal of Policy Analysis and Management, 27*(3), 457-478.

Singh, A. (2015). Private school effects in urban and rural India: Panel estimates at primary and secondary school ages. *Journal of Development Economics, 113*, 16-32.

Thomas, J. L. (2012). Combination Classes and Educational Achievement. *Economics of Education Review, 31*(6), 1058-1066.

Todd, P., & Wolpin, K. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal, 113*(February), F3 - F33.

Todd, P., & Wolpin, K. (2007). The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps. *Journal of Human Capital, 1*(1), 91-136.

Walker, S., Wachs, T., Gardner, J. M., Lozoff, B., Wasserman, G., Pollitt, E., . . . Group, I. C. D. S. (2007). Child development: risk factors for adverse outcomes in developing countries. *The Lancet, Volume 369*(9556), 145-157.

Appendix 1
The risk of bias under the standard decomposition rule

Here we illustrate how the rule of grouping all child, family and household characteristics in a single category of "home influences" will introduce a bias unless there are no omitted inputs from the school environment or these omitted inputs do not respond to families' choices.

Consider a value added specification and the presence of a single unobservable input from period 2 (U_{i2}).[27] This yields the following empirical version of the production function of skill:

$$T_{i2} = \rho T_{i1} + H_{i2}'\gamma_1 + S_{i2}'\phi_1 + h_{i2}\varphi_1 + f_i{}'\lambda + [U_{i2}\vartheta + \mu_{i0}\beta + \varepsilon_{i2} - \rho\varepsilon_{i1}] \qquad (1.1)$$

The elements in brackets at the right hand side of (1.1) correspond to the error term.

The demand functions of the inputs of skill depend on predetermined household, family and child characteristics that influence skill directly (f_i) and other exogenous input determinants capturing differences in resources, prices, environments and preferences (z_i, G_i). To ease the exposition, we will consider the urban/rural group indicator (G_i) separate from the rest of input determinants. Assume that these demand functions can be expressed linearly. Thus, for the unobserved input we have:

$$U_{i2} = z_i'\delta + f_i{}'\kappa + \tau G_i + v_{i2} \qquad (1.2)$$

Where v_{i2} captures random shocks to the demand function. If we replace (1.2) in (1.1) and collect terms, it is possible to build the following linear hybrid value added specification:

[27] The analysis can be extended to the more general case were we have several omitted inputs from both periods without affecting its main results.

$$T_{i2} = \rho T_{i1} + \ H_{i2}'\gamma_1 + S_{i2}'\phi_1 + h_{i2}\varphi_1 + f_i'(\lambda + \vartheta\kappa) + z_i'\vartheta\delta + \tau\vartheta G_i + \\ [\vartheta v_{i2} + \mu_{i0}\beta + \varepsilon_{i2} - \rho\varepsilon_{i1}] \qquad (1.3)$$

Its empirical version is given by:

$$T_{i2} = \rho T_{i1} + \ H_{i2}'\gamma_1 + S_{i2}'\phi_1 + h_{i2}\varphi_1 + f_i'\tilde{\pi} + z_i'\tilde{\psi} + \tilde{\theta}G_i + e_{i2}^{VA} \qquad (1.4)$$

The difference in average outcomes observed between the urban and rural domain can be expressed as:

$$\bar{T}_{U2} - \bar{T}_{R2} = (\bar{T}_{U1} - \bar{T}_{R1})\hat{\rho} + (\bar{H}_{U2} - \bar{H}_{R2})'\hat{\gamma}_1 + (\bar{S}_{U2} - \bar{S}_{R2})'\hat{\phi}_1 + \left(\bar{h}_{U2} - \bar{h}_{R2}\right)\hat{\varphi}_1 + \\ \left(\bar{f}_U - \bar{f}_R\right)'\hat{\tilde{\pi}} + (\bar{z}_U - \bar{z}_R)'\widehat{\tilde{\psi}} + \hat{\tilde{\theta}} \qquad (1.5)$$

As explained in the main text, one of the rules usually employed in the literature to separate the contribution of home and school influences consists is assigning all observed family, child and household characteristics to the first category. Accordingly, we will define the group comprising “period 2 home and health inputs” in the following way:

$$\widehat{H}_2 = (\bar{H}_{U2} - \bar{H}_{R2})'\hat{\gamma}_1 + \left(\bar{f}_U - \bar{f}_R\right)'\hat{\tilde{\pi}} + (\bar{z}_U - \bar{z}_R)'\hat{\tilde{\psi}} + \left(\bar{h}_{U2} - \bar{h}_{R2}\right)\hat{\varphi}_1 \quad (1.6)$$

Consistent with a standard Blinder Oaxaca decomposition, the contribution of the group indicator $\left(\hat{\tilde{\theta}}\right)$ will be assigned to a category capturing the “unexplained” part of the gap. A sufficiently large sample and the parameter structure given in (1.3) allow one to write:

$$\widehat{H}_2 = (\bar{H}_{U2} - \bar{H}_{R2})'\hat{\gamma}_1 + \left(\bar{f}_U - \bar{f}_R\right)'\left(\hat{\lambda} + \hat{\vartheta}\hat{\kappa}\right) + (\bar{z}_U - \bar{z}_R)'\hat{\vartheta}\hat{\delta} + \left(\bar{h}_{U2} - \bar{h}_{R2}\right)\hat{\varphi}_1 \quad (1.7)$$

From the demand function of the omitted input is possible to write:

$$\bar{U}_{U2} - \bar{U}_{R2} = (\bar{z}_U - \bar{z}_R)'\hat{\delta} + \left(\bar{f}_U - \bar{f}_R\right)'\hat{\kappa} + \hat{\tau} \tag{1.8}$$

And combining (1.7) and (1.8) we obtain:

$$\widehat{H}_2 = (\bar{H}_{U2} - \bar{H}_{R2})'\hat{\gamma}_1 + \left(\bar{f}_U - \bar{f}_R\right)'\hat{\lambda} + \hat{\vartheta}(\bar{U}_{U2} - \bar{U}_{R2} - \hat{\tau}) + \left(\bar{h}_{U2} - \bar{h}_{R2}\right)\hat{\varphi}_1 \tag{1.9}$$

From the term $\hat{\vartheta}(\bar{U}_{U2} - \bar{U}_{R2} - \hat{\tau})$ in (1.9) is clear that, if the omitted input belongs to the school environment, the strategy described above will introduce a bias in the contribution of the "period 2 home and health inputs" category, unless this input is not affected by families' choices ($\delta = \kappa = 0$). In the likely scenario in which the omitted school input makes a positive contribution to the cognitive gap and the differences in predetermined input determinants make a positive contribution to the gap in the school input (i.e. if $\hat{\vartheta}(\bar{U}_{U2} - \bar{U}_{R2} - \hat{\tau}) > 0$), this bias will be positive.

Notice that even if the omitted input belongs to the home environment, the strategy described above will allow only a partial account of its contribution to the cognitive gap. To fully account for this contribution, one would need to include the group indicator in $\widehat{H}_2$ and discard the notion of an "unexplained" component in the Blinder-Oaxaca sense.

Appendix 2
Coefficient estimates for all the specifications considered in the analysis

VARIABLES	Original sample & "full information"			Original sample & excluding school inputs		Complete sample & excluding school inputs	
	(1) Cumulative	(2) Value added	(3) Value added (IV)[a]	(4) Cumulative	(5) Value added	(6) Cumulative	(7) Value added
Real expenditure in child (learning materials and entertainment; round 2)	0.076 (0.090)	-0.022 (0.083)	-0.073 (0.098)	0.180 (0.106)	0.038 (0.087)	0.122*** (0.036)	0.034 (0.028)
Mother had antenatal visits during pregnancy (yes = 1)	0.132** (0.051)	0.121** (0.052)	0.115** (0.052)	0.136** (0.050)	0.119** (0.053)	0.193*** (0.044)	0.139*** (0.039)
Maternal response to child cry was affectionate (yes = 1)	0.077 (0.063)	0.025 (0.070)	-0.002 (0.071)	0.101 (0.062)	0.035 (0.065)	0.042 (0.043)	0.012 (0.046)
Child attended formal preschool (yes = 1)	0.049 (0.094)	0.003 (0.080)	-0.020 (0.071)	0.050 (0.121)	0.009 (0.113)	0.059 (0.058)	0.016 (0.044)
Household has books and child is encouraged to read (yes = 1)	0.210*** (0.057)	0.235*** (0.068)	0.248*** (0.070)	0.207*** (0.052)	0.240*** (0.062)	0.183*** (0.048)	0.194*** (0.048)
Household has a computer (yes = 1)	0.087 (0.059)	0.064 (0.055)	0.053 (0.056)	0.133* (0.075)	0.099 (0.072)	0.130*** (0.040)	0.082** (0.034)
Real expenditure in child (learning materials and entertainment; round 3)	0.062 (0.062)	0.036 (0.064)	0.023 (0.060)	0.077 (0.058)	0.041 (0.064)	0.059 (0.044)	0.029 (0.032)
Child receives help from parents when doing homework (yes = 1)	0.007 (0.093)	-0.041 (0.091)	-0.065 (0.091)	0.027 (0.085)	-0.023 (0.085)	0.076 (0.045)	0.061 (0.044)
Hours in a typical day the child spends playing	-0.003 (0.021)	-0.012 (0.026)	-0.016 (0.028)	0.032* (0.017)	0.015 (0.022)	0.037* (0.021)	0.021 (0.018)
Hours in a typical day the child spends sleeping	-0.032 (0.036)	-0.044 (0.038)	-0.050 (0.040)	-0.011 (0.032)	-0.028 (0.035)	-0.019 (0.024)	-0.036 (0.024)
Hours in a typical day the child spends studying	0.045 (0.045)	0.024 (0.041)	0.013 (0.039)	0.085* (0.046)	0.047 (0.046)	0.065*** (0.022)	0.017 (0.018)
Child is stunted (yes = 1; round 2)	-0.048 (0.085)	-0.002 (0.084)	0.022 (0.093)	-0.079 (0.101)	-0.025 (0.094)	-0.098 (0.066)	-0.024 (0.050)
Child is stunted (yes = 1; round 3)	-0.212 (0.123)	-0.184 (0.133)	-0.169 (0.130)	-0.226 (0.134)	-0.198 (0.143)	-0.125** (0.052)	-0.111** (0.053)

VARIABLES	**Original sample & “full information”**			**Original sample & excluding school inputs**		**Complete sample & excluding school inputs**	
	(1) Cumulative	**(2) Value added**	**(3) Value added (IV)[a]**	**(4) Cumulative**	**(5) Value added**	**(6) Cumulative**	**(7) Value added**
Hours in a typical day the child spends at school	-0.075 (0.051)	-0.070 (0.053)	-0.067 (0.050)	-0.042 (0.056)	-0.043 (0.054)	0.020 (0.037)	0.005 (0.034)
Years of schooling (basic education)	0.342*** (0.081)	0.253*** (0.075)	0.207*** (0.081)	0.326*** (0.077)	0.227*** (0.060)	0.254*** (0.044)	0.134*** (0.034)
CLIM: absence of problems is class (score 12-48)	0.009** (0.004)	0.011** (0.004)	0.012** (0.005)	--	--	--	--
INF: school has basic services (yes = 1)	0.183** (0.069)	0.044 (0.064)	-0.028 (0.099)	--	--	--	--
ACT: average curricular coverage (% of topics covered in depth)	0.521 (0.308)	0.388 (0.267)	0.319 (0.203)	--	--	--	--
ORG: teacher absenteeism (%)	-0.780* (0.380)	-0.761** (0.317)	-0.751*** (0.272)	--	--	--	--
ORG: school has a psychologist (yes = 1)	0.194** (0.079)	0.203** (0.087)	0.207** (0.088)	--	--	--	--
ORG: school is “multigrade” (yes = 1)	-0.292** (0.120)	-0.308*** (0.098)	-0.316*** (0.084)	--	--	--	--
TEA: more than 50% of teachers graduated from university (yes = 1)	0.090* (0.049)	0.012 (0.046)	-0.029 (0.065)	--	--	--	--
Child’s caregiver has higher education (yes = 1)	0.135* (0.066)	0.017 (0.056)	-0.043 (0.080)	0.169** (0.075)	0.025 (0.058)	0.194*** (0.058)	0.054 (0.048)
Caregiver’s age	0.013** (0.005)	0.008 (0.005)	0.006 (0.006)	0.012** (0.005)	0.007 (0.005)	0.009** (0.004)	0.006 (0.004)
Child is male (yes = 1)	0.028 (0.096)	0.034 (0.083)	0.037 (0.075)	0.056 (0.094)	0.062 (0.078)	0.068 (0.059)	0.060 (0.038)
Child’s mother tongue is Spanish (yes = 1)	0.341** (0.144)	0.389** (0.152)	0.415*** (0.152)	0.390** (0.145)	0.439** (0.170)	0.278** (0.105)	0.367*** (0.100)
Child’s age in months	0.014 (0.011)	0.004 (0.009)	-0.002 (0.007)	0.018 (0.011)	0.007 (0.010)	0.012 (0.007)	0.008 (0.006)

VARIABLES	Original sample & “full information”			Original sample & excluding school inputs		Complete sample & excluding school inputs	
	(1) Cumulative	(2) Value added	(3) Value added (IV)(a)	(4) Cumulative	(5) Value added	(6) Cumulative	(7) Value added
Child lives in urban area (yes = 1)	0.120 (0.113)	0.028 (0.096)	-0.020 (0.092)	0.395*** (0.076)	0.202** (0.091)	0.453*** (0.085)	0.246*** (0.083)
Average household total income	0.018 (0.024)	0.010 (0.021)	0.006 (0.021)	0.033 (0.032)	0.021 (0.027)	0.048** (0.018)	0.027* (0.014)
Average household size	0.013 (0.026)	0.013 (0.026)	0.013 (0.025)	0.006 (0.028)	0.006 (0.028)	0.009 (0.014)	0.008 (0.013)
Proportion of male siblings	-0.032 (0.175)	-0.100 (0.173)	-0.135 (0.175)	-0.137 (0.146)	-0.202 (0.138)	-0.012 (0.069)	-0.029 (0.052)
Child birth order	-0.097*** (0.026)	-0.085** (0.031)	-0.078** (0.033)	-0.101*** (0.025)	-0.088*** (0.029)	-0.076*** (0.020)	-0.051** (0.021)
Caregiver aspiration for child is university education (yes = 1)	0.058 (0.054)	0.026 (0.059)	0.010 (0.066)	0.112 (0.072)	0.065 (0.070)	0.216*** (0.034)	0.147*** (0.033)
Standardized raw PPVT score (round 2)		0.341*** (0.052)	0.517*** (0.165)		0.363*** (0.046)		0.399*** (0.030)
Constant	-1.362 (1.176)	0.465 (1.117)	1.406 (1.159)	-1.894 (1.152)	0.102 (1.172)	-1.667* (0.807)	-0.249 (0.813)
Observations	487	487	487	487	487	1,561	1,561
R-squared	0.557	0.608	0.594	0.510	0.573	0.473	0.557

(a) IV estimation uses round 2 Cognitive Developmental Assessment (CDA) test scores as instrument for round 2 PPVT test scores.
CDA coefficient in first stage = 0.268 (p<0.01).
Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Appendix 3
Consistent estimation of the contribution of an input category

Here we show that one can still obtain a consistent estimate of the contribution of a category of inputs when some of them remain unobserved and introduce a bias in the estimated effect of the observed inputs. The assumption required for identification in this case is that the information contained in the observed inputs suffices to predict the average urban/rural difference in the unobserved input.

To ease manipulation, assume there is only one early childhood influence (H_{i1}) and two school inputs (S_{i2}, U_{i2}), one of which remains unobserved (U_{i2}). As usual, we assume innate ability (μ_{i0}) is unobserved and skill is measured with random error (ε_{i2}), although we will ignore their potential correlation with observed inputs in order to focus on the biases caused by the omission of one of the school inputs. The empirical version of the production function of skill is given by:

$$T_{i2} = H_{i1}\gamma_2 + S_{i2}\phi_1 + e_{i2} \tag{3.1}$$

where $e_{i2} = U_{i2}\vartheta_1 + \mu_{i0}\beta_2 + \varepsilon_{i2}$.

The contribution of school inputs to the urban/rural gap in cognitive skill observed in period 2 can be expressed as:

$$CS = [E(S_{i2}|U) - E(S_{i2}|R)]\phi_1 + [E(U_{i2}|U) - E(U_{i2}|R)]\vartheta_1 \tag{3.2}$$

where expectations are taken conditioned on the child belonging to the urban $\big(E(\cdot\,|U)\big)$ or rural $\big(E(\cdot\,|R)\big)$ domain.

The estimation of production function parameters given in (3.1) by OLS will be biased due to the omission of U_{i2} if this school input is correlated with the other direct influences. In particular:

$$plim\begin{bmatrix}\hat{\gamma}_2\\ \hat{\phi}_1\end{bmatrix}=\begin{bmatrix}\gamma_2\\ \phi_1\end{bmatrix}+\vartheta_1\begin{bmatrix}\sigma_{HH} & \sigma_{HS}\\ \sigma_{HS} & \sigma_{SS}\end{bmatrix}^{-1}\begin{bmatrix}\sigma_{HU}\\ \sigma_{SU}\end{bmatrix} \tag{3.3}$$

Where σ_{ij} indicates the covariance between variables i and j. Therefore: $plim\,\hat{\phi}_1 = \phi_1+\vartheta_1[(\sigma_{HH}\sigma_{SU}-\sigma_{HS}\sigma_{HU})/(\sigma_{HH}\sigma_{SS}-\sigma_{HS}{}^2)]$

The term in brackets is the partial correlation coefficient between the unobserved and the observed school input. In other words, it corresponds to the probability limit of the OLS estimate of b_2 in the linear projection $U_{i2}=b_1H_{i1}+b_2S_{i2}+w_i$. This means that:

$$plim\,\hat{\phi}_1=\phi_1+\vartheta_1 b_2 \tag{3.4}$$

Let us assume that the information contained in the observed school input suffices to predict the urban/rural difference in the unobserved school input. In other words:

$$E(U_{i2}|U)-E(U_{i2}|R)=b_2[E(S_{i2}|U)-E(S_{i2}|R)] \tag{3.5}$$

Notice that it will not possible to consistently estimate the effect of the observed school input (this requires $b_2=0$). However, it is still possible to recover a consistent estimate of the contribution of school inputs using $\widehat{CS}=[\bar{S}_{U2}-\bar{S}_{R2}]\hat{\phi}_1$. In fact, combining (3.4) and the assumption expressed in (3.5) we obtain:

$$\begin{aligned}plim\,\widehat{CS}=plim[\bar{S}_{U2}-\bar{S}_{R2}]\hat{\phi}_1&=[E(S_{i2}|U)-E(S_{i2}|R)]plim\,\hat{\phi}_1\\ &=[E(S_{i2}|U)-E(S_{i2}|R)][\phi_1+b_2\vartheta_1]\\ &=[E(S_{i2}|U)-E(S_{i2}|R)]\phi_1+[E(U_{i2}|U)-E(U_{i2}|R)]\vartheta_1=CS\end{aligned} \tag{3.6}$$

Appendix 4

4.1 Testing interactions between school inputs and past cognitive achievement

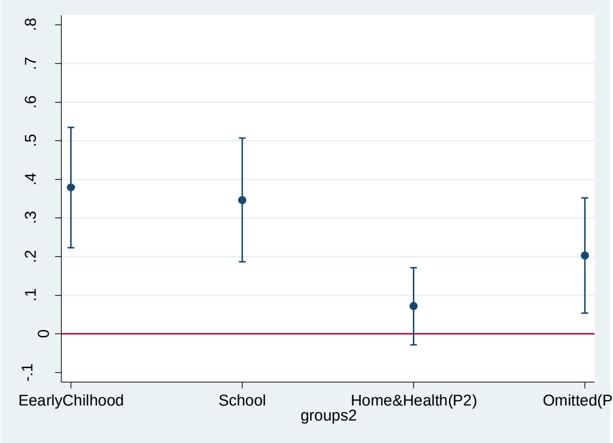
VARIABLES	Value added	Value added with interactions
Real expenditure in child (learning materials and entertainment; round 2)	-0.022 (0.083)	-0.023 (0.082)
Mother had antenatal visits during pregnancy (yes = 1)	0.121** (0.052)	0.110* (0.061)
Maternal response to child cry was affectionate (yes = 1)	0.025 (0.070)	0.034 (0.066)
Child attended formal preschool (yes = 1)	0.003 (0.080)	0.005 (0.079)
Household has books and child is encouraged to read (yes = 1)	0.235*** (0.068)	0.231*** (0.064)
Household has a computer (yes = 1)	0.064 (0.055)	0.072 (0.060)
Real expenditure in child (learning materials and entertainment; round 3)	0.036 (0.064)	0.044 (0.066)
Child receives help from parents when doing homework (yes = 1)	-0.041 (0.091)	-0.049 (0.086)
Hours in a typical day the child spends playing	-0.012 (0.026)	-0.008 (0.030)
Hours in a typical day the child spends sleeping	-0.044 (0.038)	-0.044 (0.039)
Hours in a typical day the child spends studying	0.024 (0.041)	0.022 (0.040)
Child is stunted (yes = 1; round 2)	-0.002 (0.084)	-0.006 (0.089)
Child is stunted (yes = 1; round 3)	-0.184 (0.133)	-0.172 (0.125)
Hours in a typical day the child spends at school	-0.070 (0.053)	-0.066 (0.051)
Years of schooling (basic education)	0.253*** (0.075)	0.266*** (0.075)
S1: absence of problems is class (score 12-48)	0.011** (0.004)	0.012** (0.004)
S2: school has basic services (yes = 1)	0.044 (0.064)	0.053 (0.077)
S3: average curricular coverage (% of topics covered in depth)	0.388 (0.267)	0.326 (0.227)
S4: teacher absenteeism (%)	-0.761** (0.317)	0.040 (0.446)
S5: school has a psychologist (yes = 1)	0.203** (0.087)	0.201** (0.092)
S6: school is “multigrade” (yes = 1)	-0.308*** (0.098)	-0.267* (0.124)
S7: more than 50% of teachers graduated from university (yes = 1)	0.012 (0.046)	-0.000 (0.038)
Child’s caregiver has higher education (yes = 1)	0.017 (0.056)	0.031 (0.066)

VARIABLES	Value added	Value added with interactions
Caregiver’s age	0.008 (0.005)	0.008 (0.005)
Child is male (yes = 1)	0.034 (0.083)	0.049 (0.083)
Child’s mother tongue is Spanish (yes = 1)	0.389** (0.152)	0.413** (0.155)
Child’s age in months	0.004 (0.009)	0.001 (0.009)
Child lives in urban area (yes = 1)	0.028 (0.096)	0.000 (0.092)
Average household total income	0.010 (0.021)	0.009 (0.024)
Average household size	0.013 (0.026)	0.013 (0.027)
Proportion of male siblings	-0.100 (0.173)	-0.141 (0.167)
Child birth order	-0.085** (0.031)	-0.087** (0.030)
Caregiver aspiration for child is university education (yes = 1)	0.026 (0.059)	0.026 (0.062)
Standardized raw PPVT score (round 2)	0.341*** (0.052)	0.510* (0.248)
Standardized raw PPVT score***S1**	--	-0.000 (0.005)
Standardized raw PPVT score***S2**	--	0.033 (0.066)
Standardized raw PPVT score***S3**	--	-0.403 (0.248)
Standardized raw PPVT score***S4**	--	1.107 (0.694)
Standardized raw PPVT score***S5**	--	0.031 (0.148)
Standardized raw PPVT score***S6**	--	0.094 (0.136)
Standardized raw PPVT score***S7**	--	0.035 (0.056)
Constant	0.465 (1.117)	0.702 (1.071)
Observations	487	487
R-squared	0.608	0.615
Joint significance of interactions		
F-stat	**--**	**1.16**
p-value		**0.33**

Clustered standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

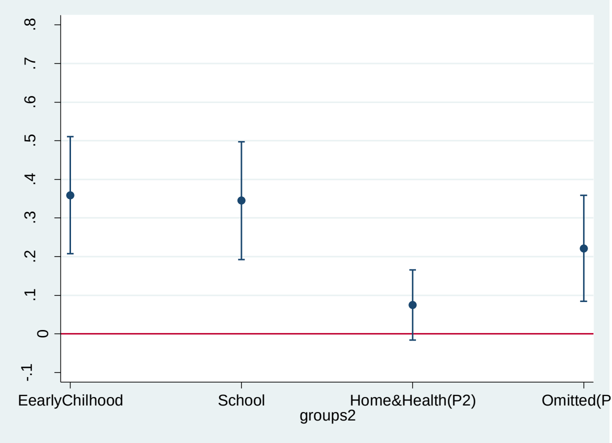
4.2 Main decomposition results allowing for interactions between input determinants (more flexible demand functions)

(A) Interactions between input determinants and the urban/rural indicator



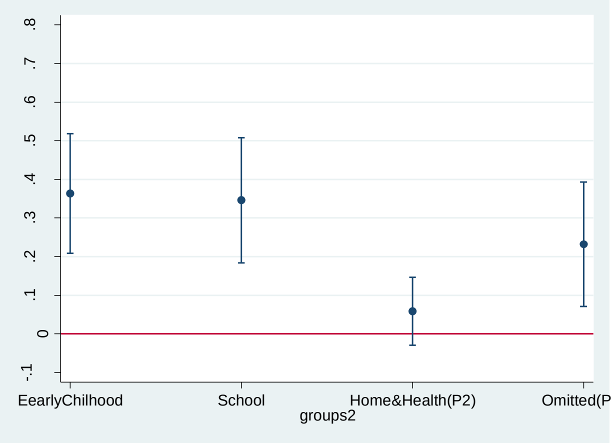
Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.10, p-value = 0.214

(B) Interactions between input determinants and the average household income



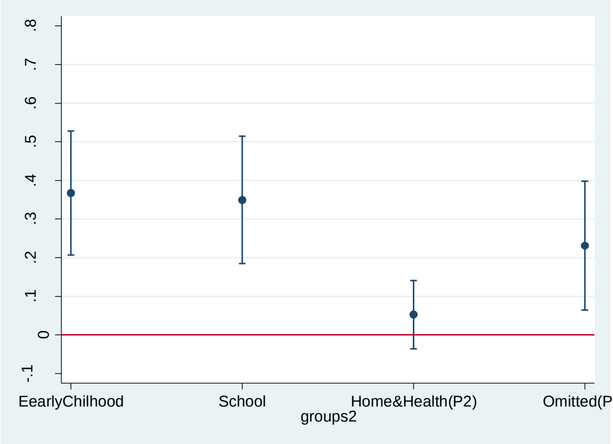
Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.11, p-value = 0.162

(C) Interactions between input determinants and the average household size



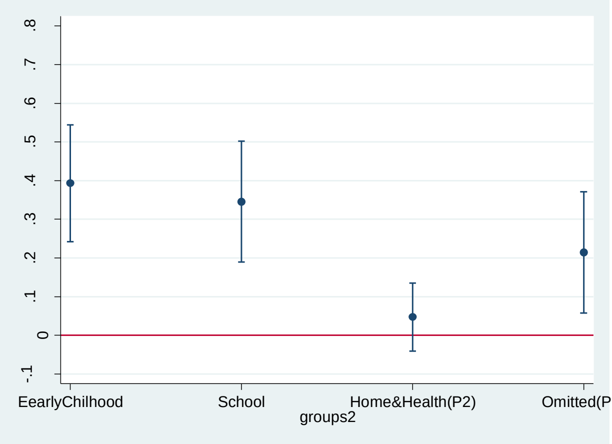
Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.13, p-value = 0.197

(D) Interactions between input determinants and the proportion of male siblings



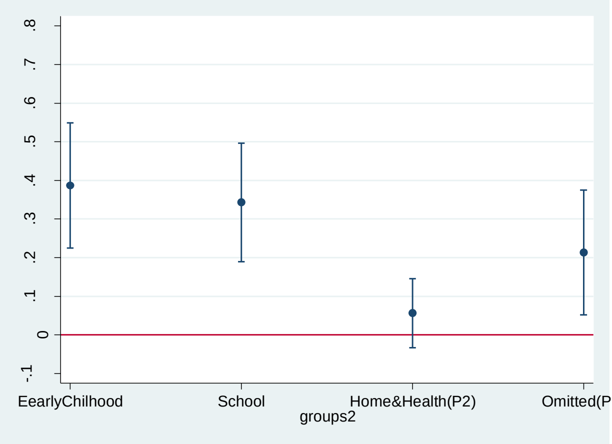
Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.12, p-value = 0.204

(E) Interactions between input determinants and the child's birth order



Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.11, p-value = 0.270

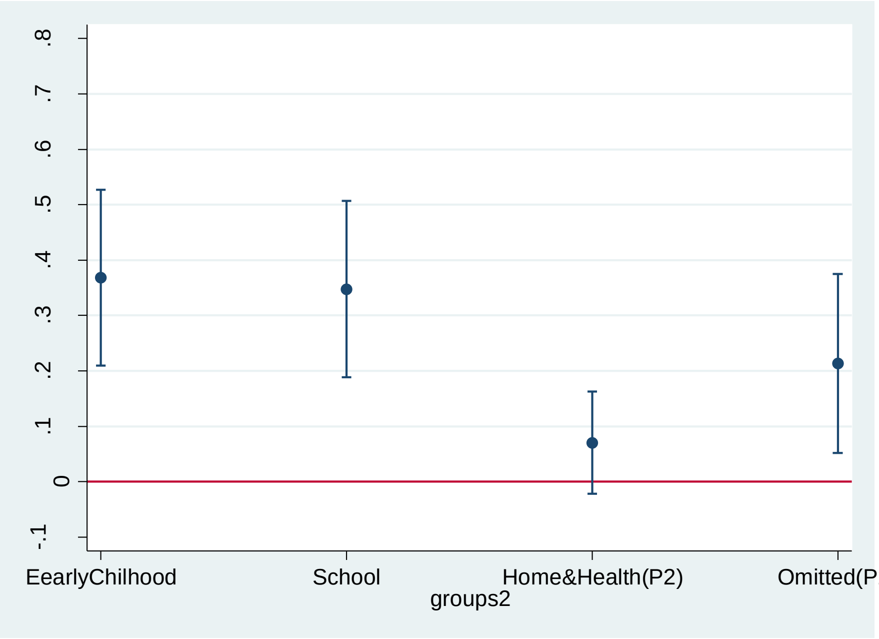
(F) Interactions between input determinants and caregivers' aspiration for child (university education)



Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.10, p-value = 0.259

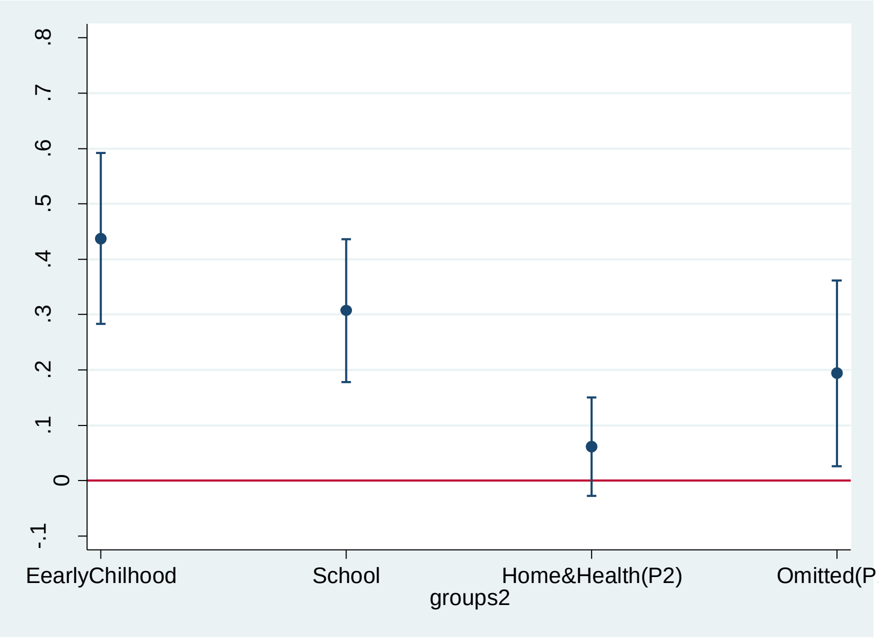
4.3 Main decomposition results with the first two principal components of the school quality dimensions

(A) Value added (original school inputs)



Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.11, p-value = 0.270

(B) Value added (first two principal components of the school quality dimensions)



Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.09, p-value = 0.352

Notes:

"EarlyChildhood" refers to early childhood influences.
"School" refers to school inputs.
"Home&Health(P2)" refers to period 2 home and health inputs.
"Omitted(P2)" refers to period 2 predetermined direct influences and period 2 omitted inputs.

4.4 Decomposition results for the cumulative specification using PPVT and Mathematics test scores

	PPVT	**Mathematics test**
(A) *"Full information" decomposition: includes all inputs from the school survey*		
Early childhood educational and health inputs	0.075 (0.058)	-0.003 (0.059)
School inputs	0.479*** (0.086)	0.456*** (0.158)
Period 2 home and health inputs	0.118*** (0.038)	0.117 (0.076)
Predetermined direct influences and omitted inputs	0.328*** (0.089)	0.430* (0.245)
(B) *Decomposition excluding inputs from the school survey (original sample)*		
Early childhood educational and health inputs	0.113 (0.071)	0.020 (0.052)
School inputs	0.062*** (0.016)	0.090*** (0.025)
Period 2 home and health inputs	0.180*** (0.042)	0.169** (0.083)
Predetermined direct influences and omitted inputs	0.645*** (0.075)	0.721*** (0.129)
(C) *Decomposition excluding inputs from the school survey (complete sample)*		
Early childhood educational and health inputs	0.096** (0.025)	0.049 (0.036)
School inputs	0.051*** (0.009)	0.125*** (0.012)
Period 2 home and health inputs	0.165*** (0.030)	0.171*** (0.0416)
Predetermined direct influences and omitted inputs	0.689*** (0.057)	0.652*** (0.091)

Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

4.5 Decomposition results for the value added specification using OLS and IV estimates

	OLS	**IV**[(a)]
Early childhood influences	0.368*** (0.081)	0.519*** (0.155)
School inputs	0.348*** (0.081)	0.280*** (0.091)
Period 2 home and health inputs	0.071 (0.047)	0.046 (0.048)
Period 2 predetermined direct influences and period 2 omitted inputs	0.213** (0.082)	0.154 (0.095)

(a) IV estimation uses round 2 Cognitive Developmental Assessment (CDA) test scores as instrument for round 2 PPVT test scores. CDA coefficient in first stage = 0.268 ($p<0.01$). Clustered standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$