

# **Linear Decompositions of Cognitive Achievement Gaps: A Cautionary Note and an Illustration Using Peruvian Data**

**Juan F. Castro**

Department of Economics - Universidad del Pacifico  
Department of International Development - University of Oxford

January 2016

## **Abstract**

Developmental gaps between children of different socioeconomic backgrounds emerge early and persist over time. Cognitive skill formation is a cumulative process and, thus, all relevant influences that took place until the time skill is measured can play role in shaping these gaps. Linear decompositions based on the Oaxaca-Blinder (OB) technique are a fairly common way of estimating the contribution of two or more categories of variables to these differences in cognitive achievement. Two prominent examples of these categories are family and school influences. In this regard, the literature shows no consensus in terms of how to implement these OB decompositions and interpret their components. It also exhibits a tendency to separate home and school influences by assigning all observed household, family and child characteristics to the first category. I argue this can lead to misleading policy implications and to biases in the estimated contributions of the categories. This analysis seeks to contribute to the literature in two ways. First, it formally explores the potential for biases in the decomposition exercises attempted so far. Second, it offers an alternative decomposition strategy consistent with explicit behavioural assumptions regarding the determination of skill inputs. This prevents arbitrary choices in terms of empirical strategy, number of components and their interpretation, and also makes the analysis less prone to biases. I illustrate empirically the main points of the paper employing a rich dataset that contains longitudinal information on cognitive test scores, family and school characteristics, to decompose the cognitive skill gap observed, at age 8, between urban and rural children in Peru.

**JEL codes:** I24, O15, C18

**Keywords:** Cognitive skill formation, Oaxaca-Blinder decomposition, Peru.

## 1. Introduction and motivation

Developmental gaps between children from disadvantaged backgrounds and those belonging to more affluent families emerge early and persist over time (Heckman, 2006, 2007; Paxson and Schady, 2007; Schady et al., 2014; Walker et al., 2007). Evidence suggests that such differences are difficult to overcome later in life, and limit these children's future economic opportunities and wellbeing (Almond and Currie, 2011; Cunha et al., 2006).

Cognitive skill formation is a cumulative process and, thus, all relevant influences that had taken place before skill is measured can, in principle, play 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 these children? Do influences originating in a particular environment (such as these children's home or school) play a major part?

The literature has tried to address this type of questions in several ways. One way has been to estimate the individual effects of particular influences, reporting their size and significance. This is the strand of literature focusing on the "technology of skill formation". The main empirical challenge related to the estimation of individual effects on skill is related to the presence of unobserved influences. Omitted influences are likely to generate biased estimates of individual effects because skill inputs are choice variables and are interrelated through the decision making process of families. This strand of the literature has explicitly exposed this problem by laying down models that postulate a production function of skill and characterize how families' choices determine its inputs (Cunha and Heckman, 2007; Glewwe and Miguel, 2008; Todd and Wolpin, 2003).

Identification of individual effects in this literature has usually relied on some form of instrumental variable strategy. This is feasible because of the limited number of parameters of interest. Findings using the US National Longitudinal Survey of the Young (NLSY/79), confirm that skill formation is a cumulative process, that socio-emotional skills affect cognitive skill, and that cognitive skill is particularly sensitive to parental investments during early childhood (Cunha and Heckman, 2008; Cunha et al., 2010; Todd and Wolpin, 2007). Efforts to replicate this in the developing world have confirmed the importance of parental investments and the fact that cognitive and socio-emotional skills are related (Helmert and Patnam, 2011; Lopez-Boo, 2009).

This strand of the literature has also made important contributions by making explicit the assumptions required by different empirical specifications to identify production function parameters (Todd and Wolpin, 2003, 2007) and by clarifying the difference between the partial and the total effect of an input on skill (Glewwe and Miguel, 2008; Todd and Wolpin, 2003)<sup>1</sup>.

---

<sup>1</sup> The partial or marginal effect of an input corresponds to its production function parameter. It implies one is holding all other direct influences constant. The total effect of an input corresponds to its partial effect plus those that occur through the changes in other inputs caused by the shift in the input of interest. Under the logic a model describing families' choices, the total effect corresponds to the parameter in a conditional demand function (see Glewwe and Miguel (2008)). It is worth noticing that experimental and quasi-experimental methods usually recover the total effect of an input.

Another strand of the literature has attempted a more direct answer to the question of which influence or group of influences is more important for the emergence of a particular gap, by proposing a linear decomposition of this gap. In particular, the difference in mean outcomes between two groups of children is decomposed into contributions that are due to the differences in the mean values of two or more categories or groups of variables. These categories are usually built in order to compare the relative importance of influences originated at home and those originated at school (Hernandez-Zavala et al., 2006; McEwan, 2004; McEwan and Marshall, 2004; McEwan and Trowbridge, 2007; Ramos et al., 2012).

Most of the studies in this strand of the literature have relied on some form of Oaxaca-Blinder decomposition technique (Blinder, 1973; Oaxaca, 1973). There are several ways to implement this technique and this implies 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 variables and contributions into different categories. In this regard, a revision of the studies applied to the developing world reveals two problematic features: (i) that the choice of decomposition strategy and interpretation of its components has been made arbitrarily (i.e. with no indication of the assumptions in terms of skill formation process and family behaviour that led to these choices); and (ii) that the rule commonly employed to separate home and school influences has been to assign all observed household, family and child characteristics to the first category.

The source of these two problematic features is the lack of a decomposition strategy based on the predictions of a framework describing the production of skill and the process determining its inputs (this is, the lack of a decomposition strategy that takes into account the insights and lessons of the “technology of skill formation” literature). The problem with these two features is that they entail the risk of producing misleading policy implications and of introducing biases in the estimated contributions of the categories of interest. This potential source of bias has been overlooked so far in the literature and emerges because several of the home and family characteristics considered within the “home influences” category can control for omitted inputs that belong to the group of “school influences”.

A good example of the above is family income or wealth. Under the logic of a model describing the production of skill and families’ choices, family income has no direct effect on skill but acts as an input determinant. In fact, family income can not only determine the quantity and quality of inputs received at home but also the quantity and quality of inputs received at school. If the latter is true, it would not be appropriate to attribute the contribution of this variable exclusively to the “family influences” category.

Based on the above, this analysis seeks to contribute to the literature by formally exploring the two problematic features of the decomposition exercises attempted so far, and by offering an alternative decomposition strategy consistent with explicit behavioural assumptions regarding the determination of skill inputs. The latter will prevent arbitrary choices in terms of decomposition technique, its components and interpretation, and will make the analysis less susceptible to biases.

For this, the rest of the paper is organized as follows. Section 2 reviews the methods and recent results associated with the decomposition exercises attempted so far for the developing world. Section 3, presents a framework describing the skill formation process and how families’ choices determine its inputs, allowing for endogenous school quality. In section 4, I use the insights of this model to formally explore the potential biases that can be introduced

by the decomposition strategies employed in the literature. I also use these insights to propose an alternative decomposition strategy less prone to these biases and to discuss its rationale under the lens of the Oaxaca-Blinder technique. In section 5, I use a rich dataset comprising cognitive test scores, family background and school information for 8-year-old Peruvian children to decompose the urban/rural gap in cognitive development, and empirically illustrate the main points made in section 4. Section 6 closes with some final remarks.

## **2. Decomposing achievement gaps in developing countries: arbitrary choices and potential biases**

Table 1 (at the end of the paper) summarizes a comprehensive list of studies that have attempted a linear decomposition of the differences in average cognitive achievement between two groups of children living in the developing world. Differences in average cognitive outcomes are expressed as a linear combination of differences in the averages of predictors, and the contributions of different subsets of predictors are estimated. These predictors or influences are typically grouped into categories that comprise family and school characteristics (see column D in Table 1).

The results of these exercises are summarized in column F of Table 3.1. A striking feature of these results is their lack of robustness regarding the contribution of school characteristics to the difference in cognitive outcomes between children of different backgrounds. In McEwan and Trowbridge (2007) and McEwan (2004), for example, the authors analysed learning outcome gaps between indigenous and non-indigenous children in Guatemala, Bolivia and Chile. They concluded that differences in the quality of schools make a significant contribution to these gaps, explaining between 50 and 70%. In Ramos et al. (2012), the authors analysed the difference in PISA results between urban and rural students in Colombia. They also found that differences in the school environment play a significant role with a contribution that ranges between 75 and 83% of the observed gap.

Results presented in Hernandez-Zavala et al. (2006) tell quite a different story. These authors addressed learning outcome gaps between indigenous and non-indigenous children in Guatemala, Mexico and Peru. They concluded that differences in “family variables” contribute more than differences in “school variables” to the overall explained gap. Surprisingly, they found that the contribution of school characteristics in Guatemala ranges between 17 and 23%, which is in sharp contrast with the results discussed in McEwan and Trowbridge (2007) where the contribution of schools to the same gap was found to be as high as 70%.

In a similar fashion, and although they do not consider an explicit category containing school variables<sup>2</sup>, in Arteaga and Glewwe (2014), the authors highlight the role played by household and child characteristics above that of community characteristics. They analysed cognitive test score gaps between indigenous and non-indigenous children in Peru and found that, by age 8 (when children are in Grade 2), differences in household and child characteristics account for 80% of the gap.

---

<sup>2</sup> The authors, however, present their community level fixed effects as partially capturing differences in school and teacher characteristics.

Further analysis of the studies reviewed in Table 1 reveals two additional characteristics that can help explain the lack of consensus regarding the importance for cognitive achievement gaps of differences in school characteristics *vis-à-vis* the importance of differences in family variables. First, the choice of decomposition strategy and the interpretation of its components (summarized in column C of Table 1) can be characterized as arbitrary. In other words, there is no explicit reference to the assumptions that have led to choosing a particular empirical strategy to decompose the observed gap and to interpret the components into which it has been decomposed.

Second, in all the studies providing an estimate of the contribution of different subsets of observed influences (e.g. family and school characteristics), control variables have been assigned to particular subsets or categories (as reported in columns D and E) without consideration of the role they play in the production of skill. The potentially harmful consequences of these two features are: (i) the risk of introducing biases in the estimate of the contribution of particular categories of variables; and (ii) the risk of overlooking the role of relevant influences when carving out policy implications. In what follows, I further develop these ideas.

## 2.1. Choosing a decomposition strategy and interpreting its components

All the studies summarized in Table 1 have used some variant of the Oaxaca-Blinder, henceforth OB, decomposition. In general, this strategy is based on decomposing the difference in mean outcomes between two groups into a portion due to differences in the mean values of observable predictors, and a portion due to differences in the coefficients governing the relationship between the outcome and these predictors. The latter is usually described as the “unexplained” part of the gap.

There are different ways to implement the OB decomposition. Depending on the number of components involved in the decomposition, these are usually classified as “twofold” or “threefold” (Jann, 2008). Let us start by exploring the “threefold” decompositions. For this, 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:

$$\begin{aligned} y_{iA} &= x'_{iA} \beta_A + \varepsilon_{iA} \\ y_{iB} &= x'_{iB} \beta_B + \varepsilon_{iB} \end{aligned} \tag{1}$$

If we estimate a linear regression for each group including an intercept in both  $x_{iA}$  and  $x_{iB}$ , the following will hold:  $\bar{y}_A = \bar{x}'_A \hat{\beta}_A$  and  $\bar{y}_B = \bar{x}'_B \hat{\beta}_B$ .

One way of measuring how much of the difference in mean outcomes has to do with differences in predictors and how much with differences in coefficients is by using the following decomposition (Jones and Kelly, 1984; Winsborough and Dickenson, 1971):

$$\bar{y}_A - \bar{y}_B = (\bar{x}_A - \bar{x}_B)' \hat{\beta}_B + \bar{x}'_B (\hat{\beta}_A - \hat{\beta}_B) + (\bar{x}_A - \bar{x}_B)' (\hat{\beta}_A - \hat{\beta}_B) \tag{2}$$

This is a “threefold” decomposition that takes group B as the reference group. The first component captures the contribution of the difference in predictors or endowments (the portion of the gap that would be closed if group B had the same endowments as group A). The second component captures the contribution of the difference in coefficients (the portion

of the gap that would be closed if group B had the same coefficients as group A). The third component is an interaction term that accounts for the fact that differences in endowments and coefficients occur simultaneously. It is the portion of the gap that only arises if endowments and returns change together (Biewen, 2012).

If the decomposition takes group A as the reference group, it yields:

$$\bar{y}_A - \bar{y}_B = (\bar{x}_A - \bar{x}_B)' \hat{\beta}_A + \bar{x}_A'(\hat{\beta}_A - \hat{\beta}_B) - (\bar{x}_A - \bar{x}_B)'(\hat{\beta}_A - \hat{\beta}_B) \quad (3)$$

The interpretation is similar to that provided in the previous paragraph but with changes occurring in the endowments and coefficients of group A.

Let us now briefly focus on the “twofold” decompositions. These are better appreciated if we introduce a third vector of reference coefficients ( $\hat{\beta}_R$ ) to measure the contributions of the differences in endowments and coefficients to the overall gap. The difference in mean outcomes between groups A and B can be expressed as follows:

$$\bar{y}_A - \bar{y}_B = (\bar{x}_A - \bar{x}_B)' \hat{\beta}_R + \bar{x}_A'(\hat{\beta}_A - \hat{\beta}_R) + \bar{x}_B'(\hat{\beta}_R - \hat{\beta}_B) \quad (4)$$

The four types of “twofold” decompositions usually encountered in the literature emerge depending on the choice of  $\hat{\beta}_R$  (see Table 2).

**Table 2**  
**Four different two-fold Oaxaca-Blinder decompositions**

	<b>Reference coefficients</b>	<b>Decomposition</b>
1	Group A coefficients ( $\hat{\beta}_R = \hat{\beta}_A$ )	$\bar{y}_A - \bar{y}_B = (\bar{x}_A - \bar{x}_B)' \hat{\beta}_A + \bar{x}_B'(\hat{\beta}_A - \hat{\beta}_B)$
2	Group B coefficients ( $\hat{\beta}_R = \hat{\beta}_B$ )	$\bar{y}_A - \bar{y}_B = (\bar{x}_A - \bar{x}_B)' \hat{\beta}_B + \bar{x}_A'(\hat{\beta}_A - \hat{\beta}_B)$
3	Coefficients from a pooled regression over both groups ( $\hat{\beta}_R = \hat{\beta}_{Pooled}$ )	$\bar{y}_A - \bar{y}_B = (\bar{x}_A - \bar{x}_B)' \hat{\beta}_{Pooled} + \bar{x}_A'(\hat{\beta}_A - \hat{\beta}_{Pooled}) + \bar{x}_B'(\hat{\beta}_{Pooled} - \hat{\beta}_B)$
4	Coefficients from a pooled regression over both groups including a group indicator ( $\hat{\beta}_R = \hat{\beta}_{Pooled*}$ )	$\bar{y}_A - \bar{y}_B = (\bar{x}_A - \bar{x}_B)' \hat{\beta}_{Pooled*} + \bar{x}_A'(\hat{\beta}_A - \hat{\beta}_{Pooled*}) + \bar{x}_B'(\hat{\beta}_{Pooled*} - \hat{\beta}_B)$

In all four cases, the first term in the right hand side of the decomposition equation captures the portion of the gap that can be explained by the differences in endowments. The remaining term(s) capture the portion of the gap that is due to differences in coefficients or the “unexplained” part of the gap.

The first two “twofold” decompositions correspond to the original formulations proposed in Oaxaca (1973) and Blinder (1973). The third decomposition was considered in Neumark (1988). This author analysed the implications in terms of firm behaviour of choosing different reference groups when estimating the contribution of “discrimination” to a wage gap. He proposed using the coefficients of a pooled regression as a reference as this requires weaker assumptions in terms of firms’ discriminatory behaviour.

Finally, the fourth “twofold” decomposition is also based on a vector of reference coefficients obtained from a pooled regression but including a group indicator. It is worth noticing that this is equivalent to estimating the “unexplained” part of the gap by using the coefficient of the group indicator in the pooled regression. To see this, consider this pooled\* regression to be as follows:

$$y_i = x_i' \beta_{Pooled*} + \delta D_i + \varepsilon_i \quad (5)$$

where  $D_i$  is the group indicator (a dummy variable that 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 the regression line passes through the means of both groups. Therefore:  $\bar{y}_A = \bar{x}_A' \hat{\beta}_{Pooled*} + \hat{\delta}$  and  $\bar{y}_B = \bar{x}_B' \hat{\beta}_{Pooled*}$ . This, in turn, implies that:

$$\begin{aligned} \bar{y}_A - \bar{y}_B &= (\bar{x}_A - \bar{x}_B)' \hat{\beta}_{Pooled*} + \hat{\delta} \\ &= (\bar{x}_A - \bar{x}_B)' \hat{\beta}_{Pooled*} + \bar{x}_A' (\hat{\beta}_A - \hat{\beta}_{Pooled*}) + \bar{x}_B' (\hat{\beta}_{Pooled*} - \hat{\beta}_B) \end{aligned} \quad (6)$$

which means that  $\bar{x}_A' (\hat{\beta}_A - \hat{\beta}_{Pooled*}) + \bar{x}_B' (\hat{\beta}_{Pooled*} - \hat{\beta}_B) = \hat{\delta}$ .

There is no consensus regarding which is the best OB decomposition. In Biewen (2012), for example, the author advocates for the “threefold” strategy arguing that the interaction term is a constituent part of the difference in means and “it is hard to find reasons to allocate [it] either in whole or in part to either the ‘characteristics’ or the ‘returns’ effect” (Biewen, 2012; p. 12). Its interpretation, however, can be quite problematic, especially if it accounts for a substantial portion the overall gap.

Within the family of “twofold” decompositions, in Elder et al. (2010) and Jann (2008) the authors advocate for the OB pooled\* option. In both studies, the authors argue that the alternative OB pooled option tends to overstate the contribution of the “explained” component. In Elder et al. (2010) the authors further show that the “unexplained” part estimated using the OB pooled\* decomposition is usually close to the estimates provided by the more standard OB options (those where the reference coefficients are set to the coefficients of group A or B). They, therefore, propose the OB pooled\* strategy as an attractive method for obtaining single measures of the “explained” and “unexplained” portions of a gap.

This lack of consensus is manifest in the literature surveyed in Table 1. Out of the ten studies surveyed, two use the “threefold” decomposition (studies 6 and 10); two use a “twofold” decomposition taking the disadvantaged group as reference (3 and 8); one uses a “twofold” decomposition taking the advantaged group as reference (5); one uses a “twofold” decomposition and reports results taking the advantaged and disadvantaged groups as reference (4); three use a “twofold” decomposition using the coefficients of a pooled regression as reference (1,2,8), and two of them use the pooled\* option (1 and 2); and one study uses both a “twofold” decomposition taking the advantaged group as reference and one taking the coefficients of a pooled\* regression as reference (study number 7).

There are also different interpretations given to the size and significance of the “unexplained” part of the gap. For example, some studies explicitly acknowledge that the “unexplained” part of the gap can be capturing the contribution of omitted influences (Hernandez-Zavala et al., 2006; McEwan, 2004; McEwan and Marshall, 2004; McEwan and Trowbridge, 2007). Other

studies, however, implicitly assume that the skill formation process has been fully specified, and interpret the difference in coefficients capturing the “unexplained” part of the gap as literally revealing a difference in the effectiveness with which inputs are transformed into skill by the two groups under analysis (Barrera-Osorio et al., 2011; Beltran and Seinfeld, 2012; Burger, 2011; Zhang and Lee, 2011).

A common feature of all the studies presented in Table 1 is that the choice of decomposition strategy is not based on a framework that describes the production of skill and the process determining its direct influences or inputs. This leads to arbitrary choices in terms of the number of components and reference groups used to build these components, as well as to arbitrary interpretations of the results regarding their contribution to the gap under analysis. This explains the lack of consensus regarding which decomposition strategy to use and can lead to misleading policy implications.

Consider, for example, those studies that interpret the “unexplained” part of the gap as revealing a difference in the effectiveness with which inputs are transformed into skill. Without an explicit reference to a production function of skill, and in case the “unexplained” part dominates, this interpretation will lead to policy recommendations that advocate for a more efficient use of school resources instead of an increase in the provision of school inputs (see, for example, Beltran and Seinfeld (2012) and Burger (2011))<sup>3</sup>. This type of policy recommendation can be misleading if the difference in coefficients is in fact a symptom of omitted inputs that are unevenly distributed between the two groups. In this regard, the risk of omitted inputs is particularly significant in the studies surveyed in Table 1, as they all rely on cross sectional data. This means that can only account for contemporaneous influences whereas the skill formation process is cumulative, which implies that outcomes observed in a particular moment of time are a function of all influences that have occurred until that moment.

## **2.2. Building up categories and assigning contributions**

In eight out of the ten studies summarized in Table 1, the authors go beyond the “difference in endowments-difference in coefficients” dichotomy and further decompose the former into different subsets or categories of variables. These categories typically comprise family and school characteristics. The objective, thus, is to measure the contribution of influences related to these two environments to the gap under analysis. The discussion that follows is centred around empirical exercises that focus on these two categories. The main messages, however, can be generalized to situations that involve more than two categories.

Estimating the contribution of family and school influences to the gap under analysis requires an estimate of the effects of these influences on skill and a rule to assign these influences into the categories proposed. Both elements entail the risk of introducing a bias in the estimate of the contributions of family and school influences. The first source of bias has been widely addressed in the literature focused on the “technology of skill formation” and is related to the presence of omitted variable biases in the estimates of production function parameters.

---

<sup>3</sup> It is also worth noticing that this interpretation is consistent with the notion that the learning process is an attribute of the school and not an attribute of the child. Conceptualizing the skill formation technology as an attribute of the school implies that the learning process would cease in absence of the school, just as production would stop in absence of the firm. This implication is especially problematic when modelling broad forms of skill whose acquisition is a process that started before and continues beyond the schooling period.

Because we seldom observe all the relevant direct influences of skill and these are interrelated through the decision making process of families, it is highly likely that omitted influences will produce biased estimates of the direct effects of those influences we observe. Different empirical specifications of the production function of skill require different assumptions to be able to recover these parameters. These assumptions have been discussed at length in Todd and Wolpin (2003) and Todd and Wolpin (2007)<sup>4</sup>.

The second source of bias has not been addressed yet in the literature and is related to the use of rules (explicit or implicit) that end up assigning the contribution of variables that belong to one category to another. The problem arises when assigning the contribution of variables that control (directly or indirectly) for omitted influences that belong to more than one category. If this is the case, the contribution of these controls should not be assigned exclusively to either the family or school environment. If omitted influences have a positive effect on skill and the gap in their endowment is also positive, doing so will lead to overstating the contribution of the category hosting these controls.

Studies surveyed in Table 1 exhibit the abovementioned problem in two different ways. The first has to do with the assignment of predetermined household, family and child characteristics to the “family influences” category. The second is related to the use of school fixed effects or school-level averages of child characteristics and their assignment to the “school influences” category.

The inclusion of predetermined child, household and family characteristics that do not have a direct effect on skill (such as family income, household size or the child’s birth order) in the estimation of a production function is justified insofar they are relevant arguments in the demand functions of omitted direct influences. Given constraints and preferences, parents play a major role deciding the inputs that determine the skill formation process of their children. Because of this, arguments in the demand function of inputs are related to child, family and household characteristics. Inclusion of these predetermined controls implies we are replacing the omitted influences by their corresponding demand functions. This configures what is known as a “hybrid” specification (Rosenzweig and Schultz, 1983; Todd and Wolpin, 2007).

As shown in Table 1, a rule commonly employed in the literature when assigning variables into categories has been to group all family, household and child characteristics into the “family influences” category. A quick revision of the variables typically considered within this category reveals that these include direct influences (such as books or time that parents spend with children) but also variables that reflect family resources and preferences which are, therefore, controlling for omitted influences.

---

<sup>4</sup> In Fortin et al. (2011), the authors claim that correlation between the error term and covariates can still allow one to obtain consistent estimates of the “unexplained” part of a decomposition, as long as the dependence structure is the same in the two groups under analysis (what they refer to as the ignorability assumption). A decomposition that goes beyond separating the gap between an “explained” and an “unexplained” component, however, will require stronger assumptions. In this regard, Castro and Rolleston (2015) discuss how the omission of a relevant input of skill can still allow one to recover a consistent estimate of the contribution of a category of variables as long as the effect of the omitted input is picked up by observed influences that belong to its same category.

Family resources and preferences determine the inputs provided in the home environment (such as early stimulation opportunities or learning material) but can also play a role determining the quantity and quality of inputs provided at school through parents' school choices. More affluent families can provide better stimulation opportunities to their children during early childhood and can also afford enrolling them in better schools. Because of this, the rule employed in the literature entails the risk of overstating the relative importance influences provided in the family environment. This risk grows larger as families' school choices have a greater influence on the quality of school inputs and as less information on school inputs is available for the analysis.

This type of bias is likely affecting the results discussed in McEwan and Marshall (2004) and Hernandez-Zavala et al. (2006), where the authors found that family variables contribute more than schools to the gap under analysis, after grouping all family, household and child characteristics into a single category. In addition, it is worth noting that in Hernandez-Zavala et al. (2006), school data was especially limited for Mexico and the contribution of school influences was found to be zero (see Table 1, row 5, columns E and F).

Not all the studies that have grouped family and household characteristics into a single category have found that schools have a limited contribution. In fact, three studies that found that differences in school characteristics play a major role also followed this rule (McEwan, 2004; McEwan and Trowbridge, 2007; Ramos et al., 2012) (see rows 1, 2 and 3 in Table 3.1). A common feature of these studies that can explain this result is that they have captured school influences by introducing school fixed effects or the mean socioeconomic level of the peer group.

In principle, one could argue that the contribution of school fixed effects or school-level averages of child characteristics belong to the "school characteristics" category. School fixed effects absorb all direct influences that are invariant within schools and school inputs are surely among these. However, influences originated at school might not be the only inputs shared by students that belong to the same school. The stronger the correlation between children's socioeconomic status and the quality of schooling received, the closer the match between children's early childhood skill and school choice. Under this setting, poor information on early childhood inputs or past skill measures (as in the three studies mentioned above) will lead to school fixed effects or school-level averages of child characteristics absorbing omitted non-school influences, and to an overestimation of the contribution of the school environment<sup>5</sup>.

Finally, it is worth considering the strategy and results discussed in Arteaga and Glewwe (2014). These authors conclude that differences in household and child characteristics play a major role when explaining the learning outcome gap between 8-year-old indigenous and non-indigenous children in Peru. They also group all family and household characteristics into a single category, including variables that can be considered direct influences of skill and also others that belong to the demand function of omitted inputs. Different from Hernandez-Zavala et al. (2006), however, these authors do not build another category of "school

---

<sup>5</sup> Interestingly, in Ramos et al. (2012) the authors included the mean socioeconomic level of the peer group among the "school characteristics" category but in their conclusions interpreted its contribution as a family influence. This is reasonable insofar these school averages capture children's early childhood skill, but entails the same risk of understating the importance of schools as the rule of assigning all observed family and household characteristics into a single category.

characteristics” but instead measure the contribution of community-level influences captured through community fixed effects.

Community-level characteristics can exert a direct influence on the skill formation process (through interactions between the child and community members and peers) although probably its major influence (especially among young children) occurs by affecting the quantity and quality of inputs that the child receives both at home and at school. It is reasonable to postulate, therefore, that both the “household characteristics” and “community characteristics” categories analysed in Arteaga and Glewwe (2014) comprise elements that control for omitted inputs that belong to both the home and school environments.

In this case, the possibility of bias in their estimated contributions is less clear, as we can no longer say that part of the contribution of one of the categories has been assigned to the other, as in the cases discussed above. The fact that the two categories comprise elements that control for omitted inputs, however, introduces another type of complication that turns the analysis less informative for policy. In particular, it entails the risk of obscuring the role of potentially relevant inputs which, in this case, are likely related to the school environment. In other words, school inputs which are potentially relevant in explaining the gap under analysis and that can be directly affected by policy action end up subsumed under the “household characteristics” and “community characteristics” categories. As a consequence, policy recommendations end-up focusing on family characteristics less amenable to policy action such as parental education.

In a similar fashion to the arbitrary choices of decomposition strategy and interpretation of its components, the problems discussed in this section can also be traced back to the lack of an explicit framework describing the skill formation process and families’ choices determining its inputs. In particular, the lack of such framework leads one to overlook the difference between skill inputs and skill input determinants (those variables that belong to the demand function of inputs). This, in turn, increases the risk of using rules that end-up assigning the contribution of one category to another.

### **3. The production function of skill and families’ choices regarding its inputs**

In this section I describe the skill formation technology and present a simple model describing how families’ choices determine its inputs. The objective is to formalise the difference between the inputs to skill formation and the variables that determine these inputs, postulate how are they related, and describe the potential roles that input determinants can play in an empirical model seeking to explain the skill formation process. This will serve to illustrate the risk of bias if one assigns contributions to variable categories following the rule commonly employed in the literature. It will also serve to guide the design of an alternative decomposition strategy that mitigates this risk.

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.

Let us now define the production function of skill. Skill exhibited 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:

$$A_{i2} = A_2(HI_{i2}, HI_{i1}, SI_{i2}, SY_{i2}, h_{i2}, h_{i1}, f_i, \mu_{i0}) \quad (7)$$

where  $HI_{i1}$  are educational inputs provided during early childhood (period 1);  $HI_{i2}$  are educational inputs provided at home during period 2;  $SI_{i2}$  are educational inputs provided at the school where the child is enrolled during period 2;  $SY_{i2}$  are 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  $\mu_{i0}$  is the child's innate ability.

Importantly, expression (7) 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. For this analysis, I 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 arguments in this production function are similar to those proposed in Glewwe and Miguel (2008) except for the presence of  $f_i$ . This formulation, thus, allows for predetermined child and parental characteristics (e.g. parental education) to have a direct influence on skill.

The fact that inputs are choice variables and we seldom observe all relevant influences complicates the estimation of their effects due to endogeneity problems<sup>6</sup>. However, this same fact can provide important insights regarding the different types of relations that can be postulated between children's skill and its determinants. A clear understanding of these relations will play an important role in the design of a decomposition strategy that minimizes the risk of incurring in biases when building up categories and assigning contributions. For this, we first need to consider a model describing families' choices.

The model presented here follows Glewwe and Miguel (2008) closely but extends their original formulation to allow for endogenous school inputs. In Glewwe and Miguel (2008), the authors assume that school and teacher characteristics available to the child are not influenced by parental decisions made during the period under analysis (between the child's conception and the end of the primary school cycle). During this period, families' choices related to the school environment are limited to the number of years of schooling. This is consistent, for example, with a situation where school inputs are solely a function of the family's location decision and this decision was made prior to the period under analysis and cannot be changed. It should be noticed that if the location decision can fully characterize the school inputs available to the child, the supply of educational services within each locality must be fairly homogeneous.

In this regard, it is reasonable to assume that parents can influence the school and teacher characteristics available to their children either by changing location (migrating) or because

---

<sup>6</sup> In this case, "endogeneity" is understood as the presence of correlation between observed and unobserved influences.

localities are better characterized by a distribution of educational services from where parents can choose, rather than by a homogeneous type of school. Under 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, I will adopt a more flexible approach. I will assume that families can choose a particular school ( $j$ ) with a particular set of characteristics ( $SI_{ij}$ ) from a given set  $S_i = \{SI_{i1}, \dots, SI_{ij}\}$ . This set is not necessarily the same for all families and is not necessarily defined by the locality where the family was settled at the beginning of the period. This allows for differences in the distribution of educational services across localities and for migration during the period under analysis. In general terms, this set is defined by the distribution of educational services available in the geographical area within which migration typically occurs, a characteristic which is specific to the context under analysis.

In the extreme case in which families do not change location during periods 1 and 2, this area will be defined by the locality where the family was established at the beginning of period 1. If families typically move across the entire territory or country under analysis, the school choice set will no longer be an additional source of heterogeneity and  $S_i = S$ ; that is, the supply of educational services available to each family will have the same characteristics.

At this point it is worth noticing that the objective of this model is not to explain how location decisions are taken and how these affect the quality of school services available to the child. The objective is to illustrate the relation between the inputs of skill and its determinants, allowing family choices to affect school characteristics considering the fact that the supply of educational services available to each family is not necessarily the same.

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

$$U_i = U(C_{i1}, C_{i2}, h_{i2}, h_{i1}, A_{i2}; \tau, \sigma, \omega) \quad (8)$$

where  $C_{it}$  is child  $i$ 's parental consumption of an aggregate good in period  $t$ , and  $\tau$ ,  $\sigma$  and  $\omega$  reflect parental preferences regarding time, child's skill and child's health, respectively.

Child health is determined according to the following production functions:

$$h_{i1} = H_1(c_{i1}, M_{i1}, HE_{i1}, \eta_{i0}) \quad (9)$$

$$h_{i2} = H_2(h_{i1}, c_{i2}, M_{i2}, HE_{i2}, \eta_{i0}) \quad (10)$$

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

Under this setting, parents choose consumption levels ( $C_{it}$  and  $c_{it}$ ), health inputs ( $M_{it}$ ), educational inputs provided during early childhood and at home ( $HI_{i1}$ ,  $HI_{i2}$ ), and years of schooling in a particular school ( $SY_{i2}$ ) to maximize utility given in (8), subject to the skill

formation technology given in (7), the production functions for health given in (9) and (10), and the following budget constraint:

$$Y_{i1} - S_{i1} = p_{c1}(C_{i1} + c_{i1}) + p_{m1}M_{i1} + p_{h1}HI_{i1} \quad (11)$$

$$Y_{i2} + (1 + r)S_{i1} = p_{c2}(C_{i2} + c_{i2}) + p_{m2}M_{i2} + p_{h2}HI_{i2} + \sum_{j=1}^{J_i} p_s^j SY_{i2}^j 1(SY_i^j = SY_{i2}) \quad (12)$$

In (11) and (12),  $S_{i1}$  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  $t$ ,  $p_{h1}$  is the price of educational inputs provided during early childhood,  $p_{h2}$  is the price of educational inputs provided at home during period 2,  $p_s^j$  is the price of one year of schooling at school  $j$ ,  $SY_i^j$  is the number of years of schooling demanded at school  $j$ , and  $1(SY_i^j = SY_{i2})$  is an indicator function that equals 1 in case school  $j$  has been chosen ( $SY_i^j = SY_{i2}$ ) and 0 otherwise<sup>7</sup>. Finally,  $Y_{it}$  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, I assume that parents can choose a particular school ( $j$ ) from a given set ( $S_i$ ;  $|S_i| = J_i$ ). This feature of the model implies that parents will be able not only to choose the number of years of schooling, but that they can also influence the educational inputs provided at school. In fact, by choosing a certain number of years of schooling at a particular school, parents are also determining that their child will be exposed to a certain quality of educational inputs. This means that we need an additional expression to fully characterize the optimization problem faced by parents. Formally:

$$SI_{i2} = \sum_{j=1}^{J_i} SI_{ij} 1(SY_i^j = SY_{i2}) \quad (13)$$

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 ( $r, p, p_s^j$ )  $j = 1, \dots, J_i$  and  $p = (p_{c1}, p_{c2}, p_{m1}, p_{m2}, p_{h1}, p_{h2})$ ; (iii) exogenous environmental variables ( $HE_{i1}, HE_{i2}, S_i$ ); (iv) predetermined direct influences ( $f_i$ ); (v) endowments ( $\mu_{i0}, \eta_{i0}$ ); and (vi) preferences ( $\tau, \sigma, \omega$ ). Formally:

$$C_{it}^* = C_t(Y_{i1}, Y_{i2}; r, p, p_s^j; HE_{i1}, HE_{i2}, S_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega) \quad t = 1, 2; j = 1, \dots, J_i \quad (14)$$

$$c_{it}^* = c_t(Y_{i1}, Y_{i2}; r, p, p_s^j; HE_{i1}, HE_{i2}, S_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega) \quad t = 1, 2; j = 1, \dots, J_i \quad (15)$$

$$M_{it}^* = M_t(Y_{i1}, Y_{i2}; r, p, p_s^j; HE_{i1}, HE_{i2}, S_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega) \quad t = 1, 2; j = 1, \dots, J_i \quad (16)$$

<sup>7</sup> Notice I am assuming that children do not switch schools during the period under analysis.

$$HI_{it}^* = H_t(Y_{i1}, Y_{i2}; r, p, p_s^j; HE_{i1}, HE_{i2}, S_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega) \quad t = 1, 2; j = 1, \dots, J_i \quad (17)$$

$$SY_{i2}^* = SY(Y_{i1}, Y_{i2}; r, p, p_s^j; HE_{i1}, HE_{i2}, S_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega) \quad j = 1, \dots, J_i \quad (18)$$

$$SI_{i2}^* = \sum_{j=1}^{J_i} SI_{ij} 1(SY_i^j = SY_{i2}^*) \quad (19)$$

The production function indicated in (7) 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 the hybrid production function already mentioned in section 2. In what follows I briefly describe these functions to clarify the role that exogenous input determinants can play in the estimation of the production function of skill<sup>8</sup>.

The demand function involves only predetermined variables that can have a direct or indirect effect on skill. It can be obtained by replacing (17), (18) and (19) in (7), replacing (15) and (16) in (9) and (10) and solving the demand for child's health in periods 1 and 2, and inserting these solutions into (7). This yields:

$$A_{i2} = A_2^D(Y_{i1}, Y_{i2}; r, p, p_s^j; HE_{i1}, HE_{i2}, S_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega) \quad j = 1, \dots, J_i \quad (20)$$

The conditional skill demand function conditioned over input  $k$ , only involves input  $k$  and controls for the exogenous determinants of those inputs not included. To obtain this relation we need first to consider the 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, 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:

$$HI_{it}^{CD} = H_t(SI_{i2}, SY_{i2}; Y_{CD}; p; HE_{i1}, HE_{i2}; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega) \quad t = 1, 2 \quad (21)$$

where  $Y_{CD}$  refers to resources after adjusting for school expenditures  $Y_{CD} = Y_{i1} + \frac{Y_{i2}}{1+r} - \frac{\sum_{j=1}^{J_i} p_s^j SY_{i2} 1(SY_i^j = SY_{i2})}{1+r}$ . As already noted, the price of schooling ( $p_s^j$ ) is no longer present in (21).

<sup>8</sup> For a more complete description of how to obtain and interpret these functions, the reader can consult Glewwe and Miguel (2008).

Similar expressions can be obtained for the demand for child's health in both periods after building conditional demand functions for child's consumption and health inputs. Replacing conditional demand functions for early childhood and educational home inputs and child's health in the production function given in (7) yields the demand for child's skill conditioned over school inputs. Formally:

$$A_{i2} = A_2^{CD}(SI_{i2}, SY_{i2}; Y_{CD}; p; HE_{i1}, HE_{i2}; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega) \quad (22)$$

Finally, and following the example centred on school inputs, a hybrid production function can be obtained if we replace all inputs in (7), except those related to the school environment, by their respective demand functions. Thus we obtain:

$$A_{i2} = A_2^H(SI_{i2}, SY_{i2}; Y_{i1}, Y_{i2}; r, p, p_s^j; HE_{i1}, HE_{i2}, S_i; f_i; \mu_{i0}, \eta_{i0}; \tau, \sigma, \omega) \quad j = 1, \dots, J_i \quad (23)$$

The differences between expressions (7), (21), (22) and (23) have important consequences for empirical work. Consider, for example, the difference between the effects of school inputs provided by equations (7) and (22). The effect of school inputs captured in equation (7) corresponds only to the direct impact of these inputs on skill, holding all other direct influences constant. The effect of school inputs provided by expression (22) includes this direct influence but also captures the indirect effect produced through changes in other inputs within the choice set of parents. Accordingly, experimental designs and instrumental variable techniques will typically identify the latter; i.e. they will typically identify the parameters of a conditional demand function (or the "policy effects" as denoted in Todd and Wolpin (2003))<sup>9</sup>.

A hybrid function such as the one given in expression (23) allows one to recover the parameters of the production function of observed inputs. The motivation for this type of specification is empirical and stems from the possibility of evading omitted variable biases originated by the presence of unobserved inputs. Under the rationale of a model of family choice such as the one described above, the use of exogenous input determinants in the estimation of a production function implies that the researcher believes in the possibility of omitted inputs and these have been replaced by their corresponding demand functions.

#### **4. The cognitive skill gap, empirical specifications and decomposition strategies**

In this section I use the insights provided by the model described above to illustrate the potential biases that can be introduced by the rule commonly employed in the literature to assign the contribution of individual variables to the categories proposed for the decomposition. For this, I will consider the decomposition of the cognitive skill gap observed, in period 2, between children belonging to two generic groups (A and B). I will consider two different specifications and the feasible components related to each of them. I will start discussing the potential biases present under the rule commonly employed in the literature (henceforth, the "standard decomposition rule"). I will then present an alternative decomposition strategy less prone to these biases, and discuss its rationale under the lens of the Oaxaca-Blinder technique.

---

<sup>9</sup> Notice that in an experimental setting, introducing exogenous variation in a certain input, nothing prevents post-treatment values of other inputs to change in response to treatment.

#### 4.1. The risk of bias under the standard decomposition rule

Let us assume that the production function given in (7) is approximately linear. To ease the exposition, also assume that parameters are age invariant (they only depend on the relative separation between the timing of the input and the measurement of skill) and that years of schooling ( $SY_{i2}$ ) are contained within the vector of school inputs ( $SI_{i2}$ )<sup>10</sup>. This allows one to express the production function of skill as follows:

$$A_{i2} = HI'_{i2}\gamma_1 + HI'_{i1}\gamma_2 + SI'_{i2}\phi_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f'_i\lambda_{(2)} + \mu_{i0}\beta_{(2)} \quad (24)$$

This can be regarded as a “cumulative” model where skill exhibited 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 lagged skill and period 2 influences only. For this, consider that period 1 skill can be written as:

$$A_{i1} = HI'_{i1}\gamma_1 + h_{i1}\varphi_1 + f'_i\lambda + \mu_{i0}\beta \quad (25)$$

The assumption of age-invariant parameters implies that  $\gamma_1$  and  $\varphi_1$  are the same in (24) and (25). In (25) they indicate the effect of period 1 educational and health inputs on period 1 skill. Also notice that parameters  $\lambda$  and  $\beta$  in (25) indicate the effect of predetermined direct influences and innate ability in period 1, respectively, while parameters  $\lambda_{(2)}$  and  $\beta_{(2)}$  in (24) express the cumulative effect (until period 2) of this same pair of influences.

If we subtract  $\rho A_{i1}$  from (24) and assume that the effect of inputs decays at a rate  $\rho$  we obtain<sup>11</sup>:

$$A_{i2} = \rho A_{i1} + HI'_{i2}\gamma_1 + SI'_{i2}\phi_1 + h_{i2}\varphi_1 + f'_i\lambda + \mu_{i0}\beta \quad (26)$$

where  $\lambda = \lambda_{(2)} - \rho\lambda$  and  $\beta = \beta_{(2)} - \rho\beta$ . These parameters capture the contemporaneous (period 2) effect of predetermined direct influences and innate ability. The expression given in (26) is known as a “value added” model.

Consistent estimation of the parameters involved in (24) is problematic because we seldom observe innate ability and all relevant inputs. A value added model can allow one to partially circumvent this problem if lagged skill is a sufficient statistic for assignment mechanisms that correlate with unobservable influences (e.g. if children end up sorted into different schools according to their pre-school skill). In this regard, several recent studies reviewed in Singh (2015) have shown that value added models can provide reliable estimates of the individual effects of skill inputs. A value added model, however, will not allow one to control for omitted period 2 inputs.

<sup>10</sup> The analysis can be extended to the more general case of age-dependent parameters at the cost of complicating notation with no effect on its main results.

<sup>11</sup> Appendix 1 presents more detail regarding how this assumption ensures that the model is no longer a function of period 1 inputs. It also presents the more general case of parameters that depend on child’s age. This should help clarify why the absence of period 1 inputs in (26) does not depend on the assumption of age-invariant parameters but only on the assumption that the effect of inputs decay at a rate equal to  $\rho$ .

Hybrid models stand out as a popular empirical strategy to try to circumvent the problem of omitted inputs. As already explained, the objective is to control for omitted inputs using the arguments of their corresponding demand function. Quick inspection of Table 1 reveals that all the studies that have attempted to decompose the “explained” part of the gap have relied on some form of hybrid specification. In fact, they all control for family or household characteristics that do not have a direct effect on skill but can influence it through the acquisition of educational home or school inputs.

The risk of obtaining biased estimates of the individual effects of inputs under different empirical specifications (including the value added and hybrid models presented here) has already been addressed in the literature (Todd and Wolpin, 2003, 2007). In the example that follows I will focus on another type of bias affecting linear gap decompositions that has not been acknowledged yet in the literature. As already mentioned, this has to do with the rules employed to assign variables into different categories.

For this, let us shift to the empirical versions of (24) and (25) 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 is a single unobserved input from each period, one belonging to the early childhood environment ( $HI_{i1}^U$ ) and the other to the school environment ( $SI_{i2}^U$ )<sup>12</sup>.

This yields the following empirical version of the production function of skill:

$$T_{i2} = HI'_{i2}\gamma_1 + HI'_{i1}\gamma_2 + SI'_{i2}\phi_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f'_i\lambda_{(2)} + [HI^U_{i1}\gamma_2^U + SI^U_{i2}\phi_1^U + \mu_{i0}\beta_{(2)} + \varepsilon_{i2}] \quad (27)$$

This has a value-added representation given by:

$$T_{i2} = \rho T_{i1} + HI'_{i2}\gamma_1 + SI'_{i2}\phi_1 + h_{i2}\varphi_1 + f'_i\lambda + [SI^U_{i2}\phi_1^U + \mu_{i0}\beta + \varepsilon_{i2} - \rho\varepsilon_{i1}] \quad (28)$$

The elements in brackets at the right hand side of (27) and (28) are contained in the corresponding error terms of both specifications.

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. Assume the latter are contained in a vector ( $z_i$ ) and that demand functions can be expressed linearly. Accordingly, the demand functions for omitted inputs can be written as follows:

$$\begin{aligned} HI^U_{i1} &= z'_i\delta_1 + f'_i\kappa_1 + \tau_1 G_i + v_{i1} \\ SI^U_{i2} &= z'_i\delta_2 + f'_i\kappa_2 + \tau_2 G_i + v_{i2} \end{aligned} \quad (29)$$

<sup>12</sup> The analysis can be extended to the more general case were we have several omitted inputs from both periods without affecting its main results.

Where  $v_{i1}$  and  $v_{i2}$  capture random shocks to the demand functions. Variable  $G_i$  denotes membership to the groups considered to define the gap in cognitive skill ( $G_i = 1$  if the child belongs to group A and  $G_i = 0$  if she belongs to group B). This indicator will typically have a role within  $z_i$  (and even  $f_i$ ) as achievement gaps are usually defined in terms of children's ethnicity or their geographical domain (see Table 1). In other words, I am considering the fairly general case where the group indicator defining the achievement gap can be included among the arguments of the demand function of inputs<sup>13</sup>. In (29) I have considered this variable separate from  $z_i$  and  $f_i$  to ease the exposition of the decomposition strategies.

If we replace (29) in (27) and collect terms, it is possible to build the following linear hybrid specification.

$$T_{i2} = HI'_{i2}\gamma_1 + HI'_{i1}\gamma_2 + SI'_{i2}\phi_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f'_i(\lambda_{(2)} + \gamma_2^U\kappa_1 + \phi_1^U\kappa_2) + z'_i(\gamma_2^U\delta_1 + \phi_1^U\delta_2) + (\tau_1\gamma_2^U + \tau_2\phi_1^U)G_i + [v_{i2}\phi_1^U + v_{i1}\gamma_2^U + \mu_{i0}\beta_{(2)} + \varepsilon_{i2}] \quad (30)$$

A similar exercise for the value added specification yields:

$$T_{i2} = \rho T_{i1} + HI'_{i2}\gamma_1 + SI'_{i2}\phi_1 + h_{i2}\varphi_1 + f'_i(\lambda + \phi_1^U\kappa_2) + z'_i\phi_1^U\delta_2 + \tau_2\phi_1^U G_i + [v_{i2}\phi_1^U + \mu_{i0}\beta + \varepsilon_{i2} - \rho\varepsilon_{i1}] \quad (31)$$

Expressions (30) and (31) lead to the following empirical specifications:

$$T_{i2} = HI'_{i2}\gamma_1 + HI'_{i1}\gamma_2 + SI'_{i2}\phi_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f'_i\pi + z'_i\psi + \theta G_i + e_{i2}^H \quad (32)$$

$$T_{i2} = \rho T_{i1} + HI'_{i2}\gamma_1 + SI'_{i2}\phi_1 + h_{i2}\varphi_1 + f'_i\tilde{\pi} + z'_i\tilde{\psi} + \tilde{\theta}G_i + e_{i2}^{VA} \quad (33)$$

The risk of obtaining biased estimates of the production function parameters from an OLS estimation of (32) or (33) arises from the potential correlation between observed inputs and the elements contained in the corresponding error terms of these equations. As already mentioned, this source of bias has been widely addressed in the literature and, in what follows, I will ignore it in order to focus on the bias that can arise from the rules employed to assign variables into categories as part of a gap decomposition exercise.

Let us define the achievement gap as the difference in expected skill between children belonging to groups A and B:  $E(A_{i2}|A) - E(A_{i2}|B)$ . Its empirical counterpart is given by:  $\bar{T}_{A2} - \bar{T}_{B2}$ , where upper bars indicate sample means. The inclusion of the group indicator in (32) and (33) ensures that an OLS regression passes through the mean of both groups. Thus, for the hybrid-cumulative specification we have:

<sup>13</sup> For example, the geographical domain can be a relevant argument in a demand function insofar it controls for differences in exogenous environmental variables such as the general health status or the availability of educational services in the area ( $HE_{i1}, HE_{i2}, S_i$  in the model presented above. According to the exogenous nature of input determinants, group membership should not be part of the choices made by families during the period under analysis. Consistent with this, less than 5% of children considered in the sample employed in the simulations that follow changed their geographical domain between rounds.

$$\begin{aligned}\bar{T}_{A2} - \bar{T}_{B2} = & (\overline{HI}_{A2} - \overline{HI}_{B2})' \hat{\gamma}_1 + (\overline{HI}_{A1} - \overline{HI}_{B1})' \hat{\gamma}_2 + (\overline{SI}_{A2} - \overline{SI}_{B2})' \hat{\phi}_1 + (\bar{h}_{A2} - \bar{h}_{B2}) \hat{\phi}_1 \\ & + (\bar{h}_{A1} - \bar{h}_{B1}) \hat{\phi}_2 + (\bar{f}_A - \bar{f}_B)' \hat{\pi} + (\bar{z}_A - \bar{z}_B)' \hat{\psi} + \hat{\theta}\end{aligned}\quad (34)$$

And, for the hybrid-value added we have:

$$\begin{aligned}\bar{T}_{A2} - \bar{T}_{B2} = & (\bar{T}_{A1} - \bar{T}_{B1}) \hat{\rho} + (\overline{HI}_{A2} - \overline{HI}_{B2})' \hat{\gamma}_1 + (\overline{SI}_{A2} - \overline{SI}_{B2})' \hat{\phi}_1 + (\bar{h}_{A2} - \bar{h}_{B2}) \hat{\phi}_1 \\ & + (\bar{f}_A - \bar{f}_B)' \hat{\pi} + (\bar{z}_A - \bar{z}_B)' \hat{\psi} + \hat{\theta}\end{aligned}\quad (35)$$

If one seeks to decompose  $\bar{T}_{A2} - \bar{T}_{B2}$ , the feasible components will depend on the specification being used. Both specifications allow one to identify the contribution of observed inputs provided during period 2. The hybrid-cumulative specification also allows one to separately estimate the contribution of early childhood educational and health inputs. The hybrid-value added model, however, aggregates all early childhood influences into a single component  $((\bar{T}_{A1} - \bar{T}_{B1}) \hat{\rho})$  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.

In a similar fashion to most of the empirical work surveyed in Table 1, consider a researcher using a hybrid-cumulative specification and interested in comparing the contribution of “family influences” vs. the contribution of “school influences”. Another researcher using a hybrid-value added model seeks to compare the contribution of period 2 “family influences” vs. the contribution of “school influences”. For this, both need to devise a rule determining what constitutes a family influence. As documented in section 2, the rule commonly employed in the literature has been to assign all household, family and child characteristics (both inputs and input determinants) into this category. Consider that both researchers follow this standard rule and propose the decompositions presented in Table 3.

**Table 3**  
**Categories and variables included under the standard decomposition rule**

<i>Hybrid-cumulative</i>		<i>Hybrid-value added</i>	
Family influences ( $\hat{H}$ )	$(\overline{HI}_{A2} - \overline{HI}_{B2})' \hat{\gamma}_1$ $+ (\overline{HI}_{A1} - \overline{HI}_{B1})' \hat{\gamma}_2$ $+ (\bar{f}_A - \bar{f}_B)' \hat{\pi}$ $+ (\bar{z}_A - \bar{z}_B)' \hat{\psi}$	Period 2 family influences ( $\hat{H}_2$ )	$(\overline{HI}_{A2} - \overline{HI}_{B2})' \hat{\gamma}_1$ $+ (\bar{f}_A - \bar{f}_B)' \hat{\pi}$ $+ (\bar{z}_A - \bar{z}_B)' \hat{\psi}$
School inputs	$(\overline{SI}_{A2} - \overline{SI}_{B2})' \hat{\phi}_1$	School inputs	$(\overline{SI}_{A2} - \overline{SI}_{B2})' \hat{\phi}_1$
Health inputs	$(\bar{h}_{A2} - \bar{h}_{B2}) \hat{\phi}_1$ $+ (\bar{h}_{A1} - \bar{h}_{B1}) \hat{\phi}_2$	Period 2 health inputs	$(\bar{h}_{A2} - \bar{h}_{B2}) \hat{\phi}_1$
--	--	Past influences	$(\bar{T}_{A1} - \bar{T}_{B1}) \hat{\rho}$
Unexplained	$\hat{\theta}$	Unexplained	$\hat{\theta}$

Consistent with the standard decomposition rule, the contributions of variables contained in  $f_i$  and  $z_i$  have been assigned to the category hosting family influences as they comprise characteristics that belong to this environment. Also notice that the contribution of the group indicator ( $\hat{\theta}$  or  $\hat{\theta}$  depending on the specification) has been assigned to a component labelled “unexplained”. This ensures these decompositions are consistent with the general OB approach employed in the literature. In fact, the two researchers of this example are using an OB decomposition where the “unexplained” part of the gap corresponds to the difference in coefficients measured with respect to a reference group built using a pooled regression which includes the group indicator (the decomposition labelled OB pooled\* in Table 2).

To see the potential for bias in the two decompositions given above, let us focus on the estimated contribution of the two components related to “family influences”, starting with  $\hat{H}$ . If we consider a sufficiently large sample and the parameter structure given in (30) it is not difficult to see that:

$$\begin{aligned} \hat{H} = & (\overline{HI}_{A2} - \overline{HI}_{B2})' \hat{\gamma}_1 + (\overline{HI}_{A1} - \overline{HI}_{B1})' \hat{\gamma}_2 + (\bar{f}_A - \bar{f}_B)' (\hat{\lambda}_{(2)} + \hat{\gamma}_2^U \hat{\kappa}_1 + \hat{\phi}_1^U \hat{\kappa}_2) \\ & + (\bar{z}_A - \bar{z}_B)' (\hat{\gamma}_2^U \hat{\delta}_1 + \hat{\phi}_1^U \hat{\delta}_2) \end{aligned} \quad (36)$$

From the demand functions of omitted inputs is possible to write:

$$\begin{aligned} \overline{HI}_{A1}^U - \overline{HI}_{B1}^U &= (\bar{z}_A - \bar{z}_B)' \hat{\delta}_1 + (\bar{f}_A - \bar{f}_B)' \hat{\kappa}_1 + \hat{\tau}_1 \\ \overline{SI}_{A2}^U - \overline{SI}_{B2}^U &= (\bar{z}_A - \bar{z}_B)' \hat{\delta}_2 + (\bar{f}_A - \bar{f}_B)' \hat{\kappa}_2 + \hat{\tau}_2 \end{aligned} \quad (37)$$

Combining (36) and (37) we obtain:

$$\begin{aligned} \hat{H} = & (\overline{HI}_{A2} - \overline{HI}_{B2})' \hat{\gamma}_1 + (\overline{HI}_{A1} - \overline{HI}_{B1})' \hat{\gamma}_2 + (\bar{f}_A - \bar{f}_B)' \hat{\lambda}_{(2)} + \hat{\gamma}_2^U (\overline{HI}_{A1}^U - \overline{HI}_{B1}^U - \hat{\tau}_1) \\ & + \hat{\phi}_1^U (\overline{SI}_{A2}^U - \overline{SI}_{B2}^U - \hat{\tau}_2) \end{aligned} \quad (38)$$

The presence of bias is clear as this expression involves elements that belong to the contribution of school inputs  $(\hat{\phi}_1^U (\overline{SI}_{A2}^U - \overline{SI}_{B2}^U - \hat{\tau}_2))$ . Notice that the direct contribution of predetermined direct influences  $((\bar{f}_A - \bar{f}_B)' \hat{\lambda}_{(2)})$  belongs to the “family influences” category. The contribution that operates through the demand of omitted inputs, however, does not fully belong to the “family influences” category because of the presence of omitted school inputs.

Consider a situation where the group indicator plays only a marginal role in the demand function of omitted inputs ( $\tau_1 = \tau_2 = 0$ ). In this case, the estimated contribution of family influences ( $\hat{H}$ ) would be able to account for the omitted early childhood input but at the cost

of overstating the importance of this category as it would also be including the contribution of the omitted school input<sup>14</sup>.

Let us now analyse what lies behind the estimated contribution of period 2 family influences built using the hybrid-value added specification. A sufficiently large sample and the parameter structure given in (31) allow one to write:

$$\hat{H}_2 = (\overline{HI}_{A2} - \overline{HI}_{B2})' \hat{\gamma}_1 + (\bar{f}_A - \bar{f}_B)' (\hat{\lambda}_2 + \hat{\phi}_1^U \hat{\kappa}_2) + (\bar{z}_A - \bar{z}_B)' \hat{\phi}_1^U \hat{\delta}_2 \quad (39)$$

Combining (37) and (39) we have:

$$\hat{H}_2 = (\overline{HI}_{A2} - \overline{HI}_{B2})' \hat{\gamma}_1 + (\bar{f}_A - \bar{f}_B)' \hat{\lambda}_2 + \hat{\phi}_1^U (\overline{SI}_{A2}^U - \overline{SI}_{B2}^U - \hat{\tau}_2) \quad (40)$$

In this case, the presence of a positive bias is even clearer. According to (40), the second researcher is able to fully account for the contribution of period 2 family influences (which include the period 2 effect of predetermined direct influences:  $(\bar{f}_A - \bar{f}_B)' \hat{\lambda}_2$ ), but adds to this a positive element that corresponds to the contribution of school inputs  $(\hat{\phi}_1^U (\overline{SI}_{A2}^U - \overline{SI}_{B2}^U - \hat{\tau}_2))$ .

This analysis illustrates how, unless we are able to claim the all omitted inputs belong only to a certain category, including variables that reflect predetermined direct influences or input determinants into this category will likely lead to a positive bias in its estimated contribution. In particular, it has exposed how decompositions based on the standard rule of assigning all available household, family and child characteristics into a single category can lead to an overstatement of the importance of influences related to the home or family environment *vis-à-vis* that of school inputs. This is especially significant when data on school characteristics is not particularly rich, so omitted inputs include several influences that belong to the school environment.

Based on this, in what follows I propose an alternative decomposition strategy. This strategy acknowledges the difference and relations between skill inputs and skill input determinants. I briefly describe the strategy considering the notation used in this section and explain why it can be regarded as a special type of OB decomposition.

#### 4.2. An alternative decomposition strategy: Oaxaca-Blinder with a twist

The strategy takes into account the parameter structure behind the empirical specifications of the hybrid and hybrid-value added models. This structure stems from the fact that predetermined direct influences ( $f_i$ ), exogenous input determinants ( $z_i$ ) and the group indicator ( $G_i$ ) control for all omitted inputs through their demand equations. As described in equations (30) and (32), parameters accompanying  $f_i$ ,  $z_i$  and  $G_i$  in the hybrid specification considered for the example given above are as follows:  $\pi = \lambda_{(2)} + \gamma_2^U \kappa_1 + \phi_1^U \kappa_2$ ;  $\psi = \gamma_2^U \delta_1 + \phi_1^U \delta_2$ ; and  $\theta = \tau_1 \gamma_2^U + \tau_2 \phi_1^U$ . Parameters accompanying these same variables in the

<sup>14</sup> Notice I am assuming that all inputs have a positive effect on skill (i.e. that the parameters in the production function of skill are all positive) and that group A exhibits an advantage with respect to group B.

hybrid-value added specification are as follows (see equations (31) and (33)):  $\tilde{\pi} = \lambda_{(2)} + \phi_1^U \kappa_2$ ;  $\tilde{\psi} = \phi_1^U \delta_2$ ; and  $\tilde{\theta} = \tau_2 \phi_1^U$ .

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 impose further restrictions. In addition, strong assumptions are also required to claim that omitted inputs belong only to either the family or school environment in order to assign the contribution of all the arguments belonging to demand functions to one of these categories.

Based on the above, the decomposition strategy proposed here will assign the contribution of all variables contained in  $f_i$  and  $z_i$  and the indicator function ( $G_i$ ) into a special category hosting the contribution of predetermined direct influences and omitted inputs in general. In the particular case of the hybrid-value added plus specification, this joint contribution will consider only period 2 predetermined direct influences and period 2 omitted inputs. Table 4 summarizes the categories proposed based on the contributions given in (34) and (35).

**Table 4**  
**Categories and variables included under the alternative decomposition strategy**

<i>Hybrid-cumulative</i>		<i>Hybrid-value added</i>	
Early childhood and home inputs	$(\overline{HI}_{A2} - \overline{HI}_{B2})' \hat{\gamma}_1 + (\overline{HI}_{A1} - \overline{HI}_{B1})' \hat{\gamma}_2$	Period 2 home inputs	$(\overline{HI}_{A2} - \overline{HI}_{B2})' \hat{\gamma}_1$
Health inputs	$(\overline{h}_{A2} - \overline{h}_{B2}) \hat{\phi}_1 + (\overline{h}_{A1} - \overline{h}_{B1}) \hat{\phi}_2$	Period 2 health inputs	$(\overline{h}_{A2} - \overline{h}_{B2}) \hat{\phi}_1$
School inputs	$(\overline{SI}_{A2} - \overline{SI}_{B2})' \hat{\phi}_1$	School inputs	$(\overline{SI}_{A2} - \overline{SI}_{B2})' \hat{\phi}_1$
Predetermined direct influences and omitted inputs	$(\overline{f}_A - \overline{f}_B)' \hat{\pi} + (\overline{z}_A - \overline{z}_B)' \hat{\psi} + \hat{\theta}$	Period 2 predetermined direct influences and period 2 omitted inputs	$(\overline{f}_A - \overline{f}_B)' \hat{\pi} + (\overline{z}_A - \overline{z}_B)' \hat{\psi} + \hat{\theta}$
--	--	Past influences	$(\overline{T}_{A1} - \overline{T}_{B1}) \hat{\rho}$

The parameter structure described above also allows for a simple test for omitted inputs by analysing the significance of the contribution of exogenous input determinants in the hybrid specifications. In particular, rejection of the null hypothesis  $(\overline{z}_A - \overline{z}_B)' \psi + \theta = 0$  in the hybrid-cumulative specification implies the presence of at least one omitted input. Rejection of the null  $(\overline{z}_A - \overline{z}_B)' \tilde{\psi} + \tilde{\theta} = 0$  in the hybrid-value added specification implies the presence of at least one omitted period 2 input.

From the discussion in section 2 is clear that there are different ways to implement the OB decomposition and that there are several possible interpretations for the “unexplained” part of the gap. 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. This “unexplained” part, in turn, can be

interpreted as actually capturing a difference in returns to inputs or the presence of omitted inputs.

The decomposition strategy described here can be considered as a special case of “twofold” OB decomposition. In fact, inclusion of the group indicator  $G_i$  in the hybrid models described above ensures I am using the same coefficient estimates than those required to build the OB decomposition that uses the results of a pooled regression as a reference (OB pooled\*; see Table 2). In addition, this can be regarded as an OB decomposition where the “unexplained” part of the gap is interpreted as capturing the contribution of omitted inputs<sup>15</sup>.

The distinctive feature of this decomposition strategy with respect to those surveyed in Table 1 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. In terms of an OB approach, this prevents arbitrary choices of reference group and interpretation of the “unexplained” part of the gap. It also makes explicit the difference between inputs and input determinants which, in turn, prevents the use of rules that can introduce bias by assigning the contribution of one category to another. In fact, this is the reason why, as opposed to a more standard OB decomposition, in this strategy the group indicator is not the only variable accounting for omitted inputs. Predetermined direct influences and exogenous input determinants are also included in the category hosting omitted inputs as they all have a role as arguments in their demand functions.

This feature of the decomposition strategy is related to the message conveyed by Neumark’s analysis (Neumark, 1988). In this article, the author considered the use of different reference groups when estimating the contribution of “discrimination” to a wage gap through an OB decomposition. He traced back the choice of a certain reference group to a particular assumption regarding firms’ discriminatory behaviour<sup>16</sup>. In similar fashion, if one seeks to measure the contribution of different types of inputs to a certain cognitive skill gap by means of a particular decomposition strategy, one needs to be aware of the assumptions in terms of family behaviour that allow one to recover these contributions<sup>17</sup>. In sum, to avoid arbitrary choices in terms of components and interpretations, the choice of decomposition strategy should not be made without consideration of the underlying behavioural assumptions that allow one to identify the contributions of interest.

## 5. An illustration using Peruvian data

In what follows I will empirically illustrate the main issues discussed so far. 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 direct influences originated at home *vis-à-vis* that of school inputs, especially when one lacks rich information on school

---

<sup>15</sup> Notice this means that I am not imposing the restriction of equal coefficients in both groups when estimating the empirical specifications. What I am restricting is the interpretation of this coefficient difference (or “unexplained” part) to mean the presence of omitted inputs.

<sup>16</sup> In Neumark (1988), the author showed that choosing the advantaged group as a reference implied assuming that the firm only discriminates against the disadvantaged group (“pure discrimination”). Choosing the disadvantaged group as a reference implies that the firm only prefers the advantaged group (“pure nepotism”).

<sup>17</sup> For example, assuming that family choices play no role in determining the characteristics of schools hosting their children would allow one to assign all demand function arguments to the early childhood and home inputs category. Doing this would resemble the standard decomposition rule.

characteristics; (ii) that the use of school fixed effects to account for the contribution of school inputs can lead to an overstatement of the importance of these influences, especially when one lacks information on early childhood inputs or lagged skill and schools are highly segregated; and (iii) that the alternative decomposition strategy proposed here is less prone to biases than those employed so far in the literature.

For this, I will decompose the gap in cognitive skill observed between urban and rural 8-year-old children in Peru. I will first build a “full information” decomposition relying on an unusually rich dataset that contains abundant information on school inputs and longitudinal information on cognitive achievement. This will be based on the components of the alternative decomposition strategy as described in Table 4. The objective is to verify that the school inputs considered do make a significant contribution to the gap under analysis. I will then exclude school input information and perform three additional decompositions. The first will be based on the components of the alternative 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 3. The second will be based on the components under the standard decomposition rule and will be used to verify that this decomposition introduces a positive bias in the estimated contribution of family and household influences. The third will include school fixed effects to account for school inputs and will be compared against the “full information” decomposition to assess if the fixed effects estimation tends to overstate the importance of school inputs.

### 5.1. Data sources and variables

This analysis will employ the data contained in the Peruvian dataset of the Young Lives Study<sup>18</sup>. In particular, it will consider 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. The basic time-structure of this data is summarized in Table 5.

**Table 5**  
**Time-structure and sample sizes of the relevant YL databases**

	<i>Child and household survey</i>			<i>School Survey</i> 2011
	<i>Round 1</i> 2002	<i>Round 2</i> 2006	<i>Round 3</i> 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).

All the information was merged into a single dataset at the child level. The simulations will be based on two different samples. The first considers all children that have cognitive test

<sup>18</sup> Young Lives is an international study of childhood poverty, following 12,000 children in 4 countries (Ethiopia, India, Peru and Vietnam) over 15 years.

scores for rounds 2 and 3, and attend a school included in the school survey<sup>19</sup> (487 children in 124 schools). The second sample considers all children that have cognitive test scores for rounds 2 and 3 (1,561 children).

Following the analytical framework described in section 3, period 1 variables will correspond to influences relevant from birth and up to age 5, and period 2 variables will correspond to influences relevant between ages 5 and 8. Accordingly, period 1 variables will be provided by rounds 1 and 2, while period 2 variables will be provided by round 3. Influences captured in the school survey (collected two years after round 3) will be assumed to be the same as those present in period 2<sup>20</sup>.

The measures of cognitive achievement employed in this analysis 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). The test has a Spanish version adapted for Latin America (Dunn et al., 1986) and is the only cognitive skill measure for which the younger cohort survey presents longitudinal results. The test was applied in rounds 2 and 3, when the younger cohort children were five and eight years of age, respectively.

Rounds 1, 2 and 3 of the household and child survey contain rich information on household and caregiver characteristics. Information related to the child is also fairly comprehensive, including aspects related to her health care and health status, schooling history, and time use (round 3). Table 6 presents the variables from the child and household surveys considered as early childhood and home inputs, health inputs, predetermined direct influences, and input determinants. 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<sup>21</sup>; (iv) learning activities and materials; (v) teacher characteristics; and (vi) school responsiveness<sup>22</sup>. School variables presented in Table 6 are the ones which resulted after applying a three-step procedure to narrow down the most significant predictors of cognitive skill within each of the six school input categories described above<sup>23</sup>.

---

<sup>19</sup> The risk of selection bias due to this second condition is small. Primary school attendance in Peru is close to 100% (only 0.7% of Young Lives children were not attending school in round 3). Schools participating in the school survey were randomly selected within the four strata considered by the authors of the study (urban-private, urban-public, rural-public, rural-bilingual-public; see Guerrero et al. (2012)). Even so, I will also consider the full sample of children with complete information on cognitive skill in order to explore if results are affected by the fact of restricting the sample to those children whose school was included in the school survey.

<sup>20</sup> I am assuming that school characteristics have not changed significantly during the two year period that separates round 3 from the school survey and that the child has remained in the same school since her enrolment in first grade (at age 6) until the moment in which the school survey was collected (at age 10). Accordingly, the administrative data collected from the schools included in the survey reveals that school switching is a rare phenomenon. On average, only 2% of students enrolled in primary education changed school each year between 2009 and 2010.

<sup>21</sup> Variables in this category include teachers' perception of the relations among students and between students and teachers, and of the problems and difficulties encountered during the school year.

<sup>22</sup> Variables within this category indicate whether or not the school provides support for students lagging behind or at risk of dropping out.

<sup>23</sup> It should be noted that this analysis does not aim at identifying the effect of a particular school input or to rank school inputs in terms of their importance for cognitive skill formation. I seek a reasonably comprehensive set of school and teacher characteristics to account for the contribution of school inputs, in general, to the cognitive achievement gap. The three-step procedure can be summarized as follows: (i) pairwise correlations

**Table 6**  
**Description of the variables used in the empirical specifications**

Variable type	Variable used in empirical specifications	Database
Period 1 measured cognitive skill ( $T_{i1}$ )	Standardized raw PPVT score	Round 2
Period 2 measured cognitive skill ( $T_{i2}$ )	Standardized raw PPVT score <sup>(a)</sup>	Round 3
Early childhood educational inputs ( $HI_{i1}$ )	Real expenditure in child (learning materials and entertainment; x1,000 soles; 2006 prices in urban Lima)	Round 2
	Mother had antenatal visits during pregnancy (yes = 1)	Round 1
	Maternal response to child cry was affectionate (yes = 1) <sup>(b)</sup>	Round 1
	Child attended formal preschool (yes = 1)	Round 2
Period 2 educational home inputs ( $HI_{i2}$ )	Real expenditure in child (learning materials and entertainment; x1,000 soles; 2006 prices in urban Lima)	Round 3
	Household has books and child is encouraged to read (yes = 1)	Round 3
	Household has a computer (yes = 1)	Round 3
	Child receives help from parents when doing homework (yes = 1)	Round 3
	Hours in a typical day the child spends playing	Round 3
	Hours in a typical day the child spends sleeping	Round 3
	Hours in a typical day the child spends studying	Round 3
Period 1 health input ( $h_{i1}$ )	Child is stunted (yes = 1) <sup>(c)</sup>	Round 2
Period 2 health input ( $h_{i2}$ )	Child is stunted (yes = 1)	Round 3
School inputs ( $SI_{i2}$ )	Years of schooling (basic education)	Round 3
	Hours in a typical day the child spends at school <sup>(d)</sup>	Round 3
	CLIM: absence of problems in class (score 12-48) <sup>(e)</sup>	School survey
	INF: school has basic services (yes = 1) <sup>(f)</sup>	School survey
	ACT: average curricular coverage in maths and language (average % of topics covered in depth) <sup>(e)</sup>	School survey
	ORG: teacher absenteeism (%) <sup>(g)</sup>	School survey
	ORG: school has a psychologist (yes = 1)	School survey
	ORG: school is “multigrade” (yes = 1) <sup>(h)</sup>	School survey
	TEA: more than 50% of teachers graduated from a university (yes = 1) <sup>(e)</sup>	School survey
Predetermined direct influences ( $f_i$ )	Child’s caregiver has higher education (yes = 1)	Round 3
	Caregiver’s age	Round 3
	Child is male (yes = 1)	Round 3

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.

Variable type	Variable used in empirical specifications	Database
Exogenous input determinants ( $z_i$ )	Child's mother tongue is Spanish (yes = 1)	Round 3
	Child's age in months	Round 3
	Child lives in urban area (yes = 1)	Round 3
	Average household total income (x10,000 soles; 2006 prices in urban Lima)	Rounds 2 and 3
	Average household size	Rounds 1, 2 and 3
	Proportion of male siblings	Rounds 1, 2 and 3
	Child birth order	Rounds 1, 2 and 3
	Caregiver aspiration for child's educational attainment is university education (yes=1)	Rounds 2 and 3

- (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) As reported by maths and language teachers in charge of classes attended by Young Lives children.
- (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.

Appendix 2 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.

## 5.2. Decomposition results and discussion

In this section I present and discuss the results obtained after estimating the contributions indicated in tables 3 and 4, considering the data described in Table 6<sup>24</sup>. Following the notation used in tables 3 and 4, urban children belong to group A, and rural children to group B.

Table 7 presents the estimated contributions of each component considering: (i) the alternative decomposition strategy and all the information provided by the school survey (the "full information" decomposition; see panel A); (ii) the alternative decomposition strategy

<sup>24</sup> The hybrid-value added models were estimated including also 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  $\rho$ . Consistent with the logic of a value added specification, the contributions of included period 1 inputs were assigned to the "past influences" category.

excluding the inputs provided by the school survey<sup>25</sup> and using the same sample as in (i) (see panel B); (iii) the alternative decomposition strategy excluding the inputs provided by the school survey and using the complete sample of children (see panel C)<sup>26</sup>; (iv) the standard decomposition rule excluding the inputs provided by the school survey and using the complete sample of children (see panel D); and (v) the alternative decomposition strategy excluding the inputs provided by the school survey but using school fixed effects to account for the contribution of school inputs (see Panel E).

Figure 1 shows the same point estimates 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 3 presents coefficient estimates for the variables involved in all the specifications.

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. To account for this contribution I am reporting the estimate provided by the hybrid-value added specification. As mentioned above, several recent empirical studies have shown that this specification can provide reliable estimates of the effect of contemporaneous influences on skill, revealing that lagged test scores are a sufficient statistic for input assignment mechanisms that correlate with unobservables such as omitted past inputs and innate ability. Consistent with this, the hybrid-cumulative model, which can only control for omitted inputs but retains the full cumulate effect of unobserved innate ability, shows a somewhat larger estimated contribution for school inputs.

Results reported in panel A of Figure 1 show that exogenous input determinants in the hybrid-cumulative specification have a significant contribution (22%;  $p < 0.05$ ), indicating the presence of omitted inputs being controlled for through their demand equations. It is worth noticing that in the hybrid-value added model it is no longer possible to reject the null that input determinants are non-significant. Since in this model we are controlling for all period 1 inputs (both observed and omitted), this evidence is consistent with the absence of period 2 omitted inputs<sup>27</sup>.

---

<sup>25</sup> Ignoring the information contained in the school survey implies that the only school inputs considered are years of schooling and time spent at school.

<sup>26</sup> Notice that exclusion of school input information contained in the school survey allows one to employ the complete sample of children that register a PPVT score in rounds 2 and 3.

<sup>27</sup> Failure to reject the null  $(\bar{z}_U - \bar{z}_R)' \tilde{\psi} + \tilde{\theta} = 0$  does not directly imply the absence of omitted inputs. If predetermined direct influences are a sufficient statistic in the demand equation of omitted inputs, the null  $(\bar{z}_U - \bar{z}_R)' \tilde{\psi} = 0$  will not be rejected even under the presence of omitted inputs. In this case, however, we cannot say that predetermined direct influences are a sufficient statistic in the demand equation of omitted inputs because exogenous input determinants do have a significant contribution in the hybrid-cumulative model.

**Table 7**  
**Normalized contributions to the urban/rural gap in cognitive skill at age 8**  
**(% of urban-rural gap)**

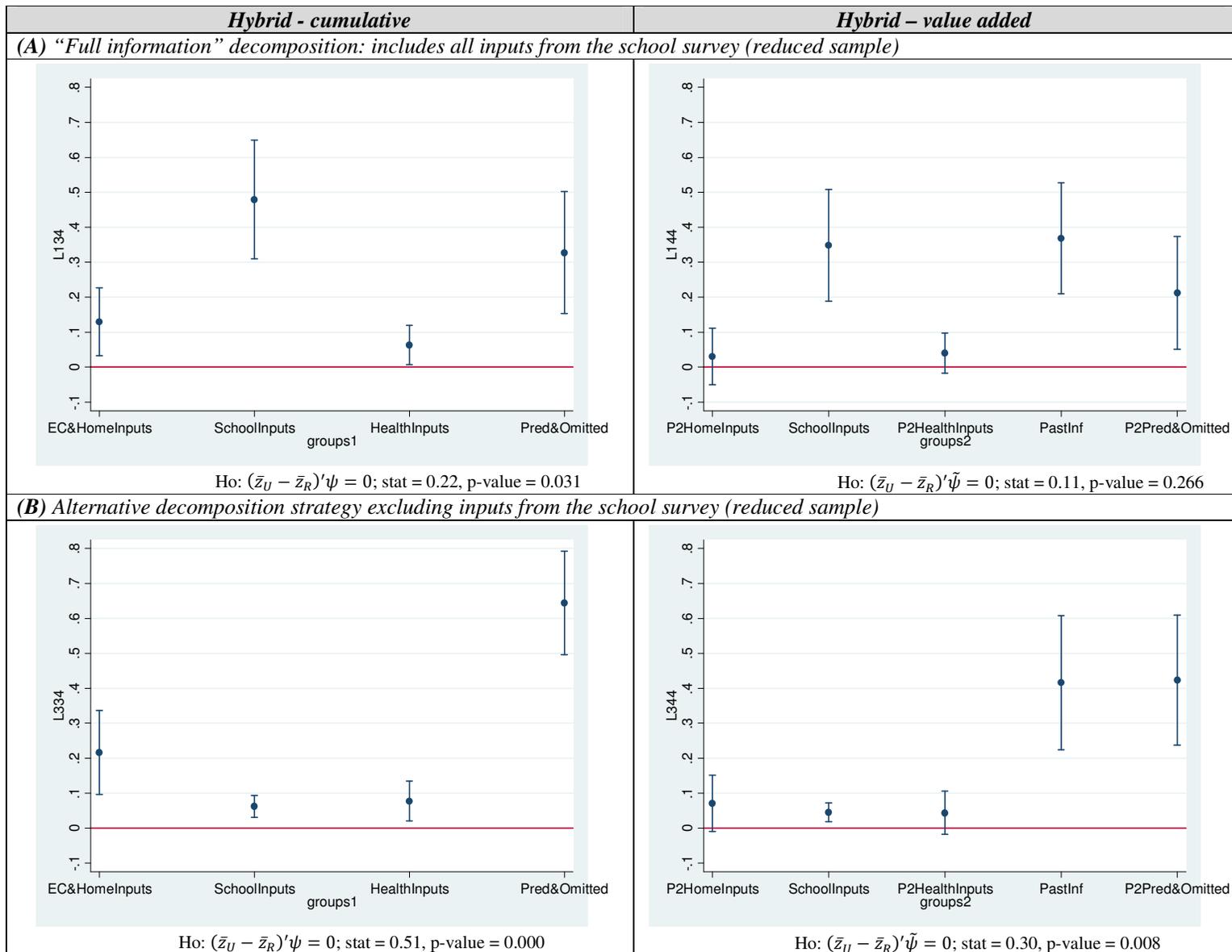
<i>Hybrid-cumulative</i>		<i>Hybrid-value added</i>	
<i>(A) "Full information" decomposition: includes all inputs from the school survey</i>			
Early childhood and home inputs	0.129*** (0.049)	Period 2 home inputs	0.030 (0.041)
School inputs	0.479*** (0.086)	School inputs	0.348*** (0.081)
Health inputs	0.063** (0.029)	Period 2 health inputs	0.041 (0.029)
--	--	Past influences	0.368*** (0.081)
Predetermined direct influences and omitted inputs	0.328*** (0.089)	Period 2 predetermined direct influences and period 2 omitted inputs	0.214** (0.082)
<i>(B) Alternative decomposition strategy excluding inputs from the school survey (reduced sample)</i>			
Early childhood and home inputs	0.216*** (0.061)	Period 2 home inputs	0.071* (0.041)
School inputs	0.062*** (0.016)	School inputs	0.045*** (0.014)
Health inputs	0.077*** (0.029)	Period 2 health inputs	0.044 (0.031)
--	--	Past influences	0.416*** (0.098)
Predetermined direct influences and omitted inputs	0.645*** (0.075)	Period 2 predetermined direct influences and period 2 omitted inputs	0.424*** (0.095)
<i>(C) Alternative decomposition strategy excluding inputs from the school survey (complete sample)</i>			
Early childhood and home inputs	0.200*** (0.024)	Period 2 home inputs	0.082*** (0.022)
School inputs	0.051*** (0.009)	School inputs	0.027*** (0.007)
Health inputs	0.060*** (0.023)	Period 2 health inputs	0.025** (0.012)
--	--	Past influences	0.419*** (0.032)
Predetermined direct influences and omitted inputs	0.689*** (0.057)	Period 2 predetermined direct influences and period 2 omitted inputs	0.447*** (0.055)

<i>Hybrid-cumulative</i>		<i>Hybrid-value added</i>	
<i>(D) Standard decomposition rule excluding inputs from the school survey (complete sample)</i>			
Family influences	0.470*** (0.046)	Period 2 family influences	0.302*** (0.045)
School inputs	0.051*** (0.009)	School inputs	0.027*** (0.007)
Health inputs	0.060** (0.023)	Period 2 health inputs	0.025** (0.012)
--	--	Past influences	0.419*** (0.032)
Unexplained	0.419*** (0.079)	Unexplained	0.228*** (0.077)
<i>(E) School fixed effects as school inputs (complete sample)</i>			
Early childhood and home inputs	0.117 (0.110)	Period 2 home inputs	0.030 (0.072)
School inputs	0.720*** (0.106)	School inputs	0.524*** (0.103)
Health inputs	0.068** (0.027)	Period 2 health inputs	0.042 (0.033)
--	--	Past influences	0.388*** (0.078)
Predetermined direct influences and omitted inputs	0.094 (0.116)	Period 2 predetermined direct influences and period 2 omitted inputs	0.016 (0.111)

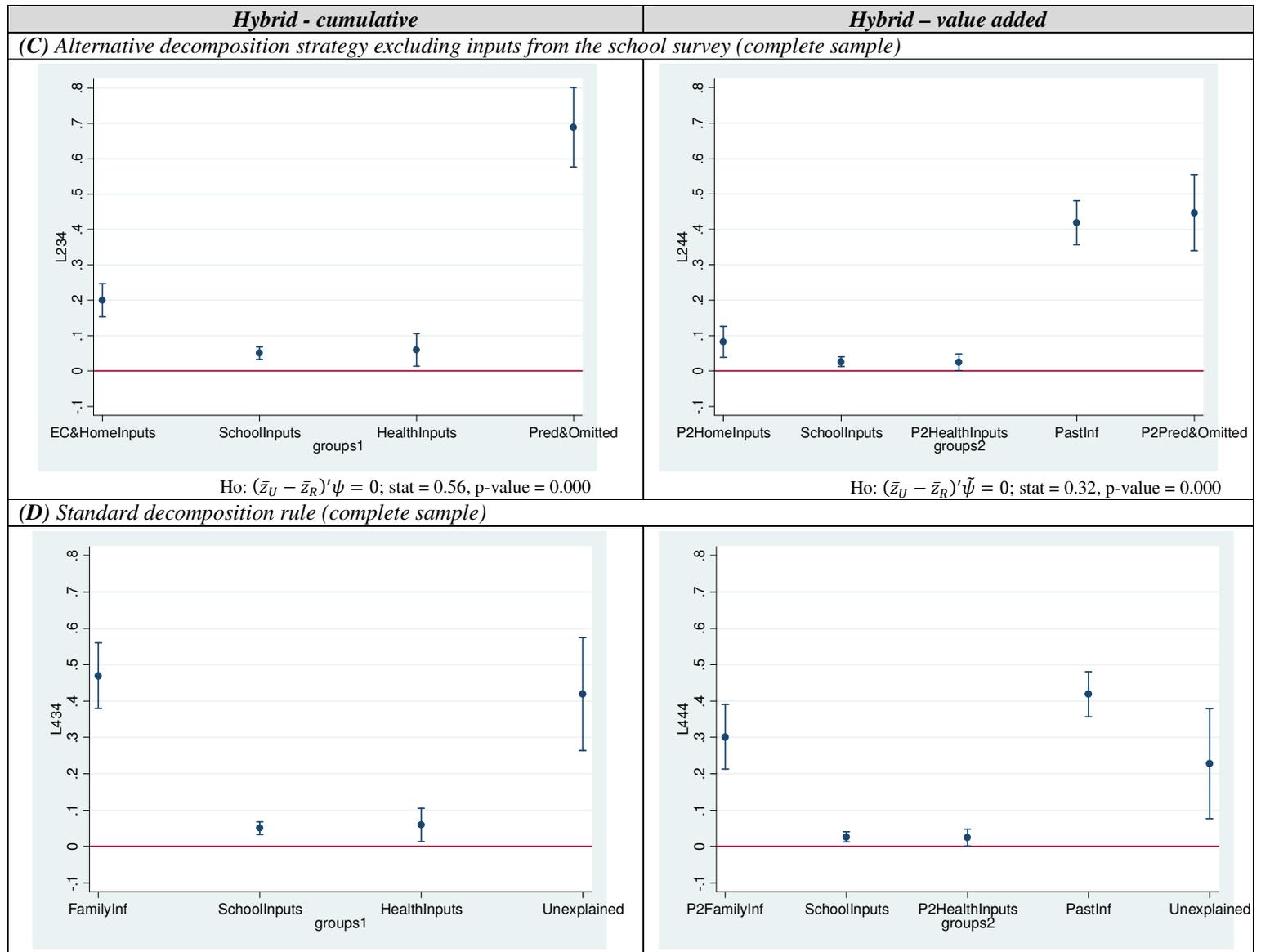
Robust standard errors in parentheses.

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

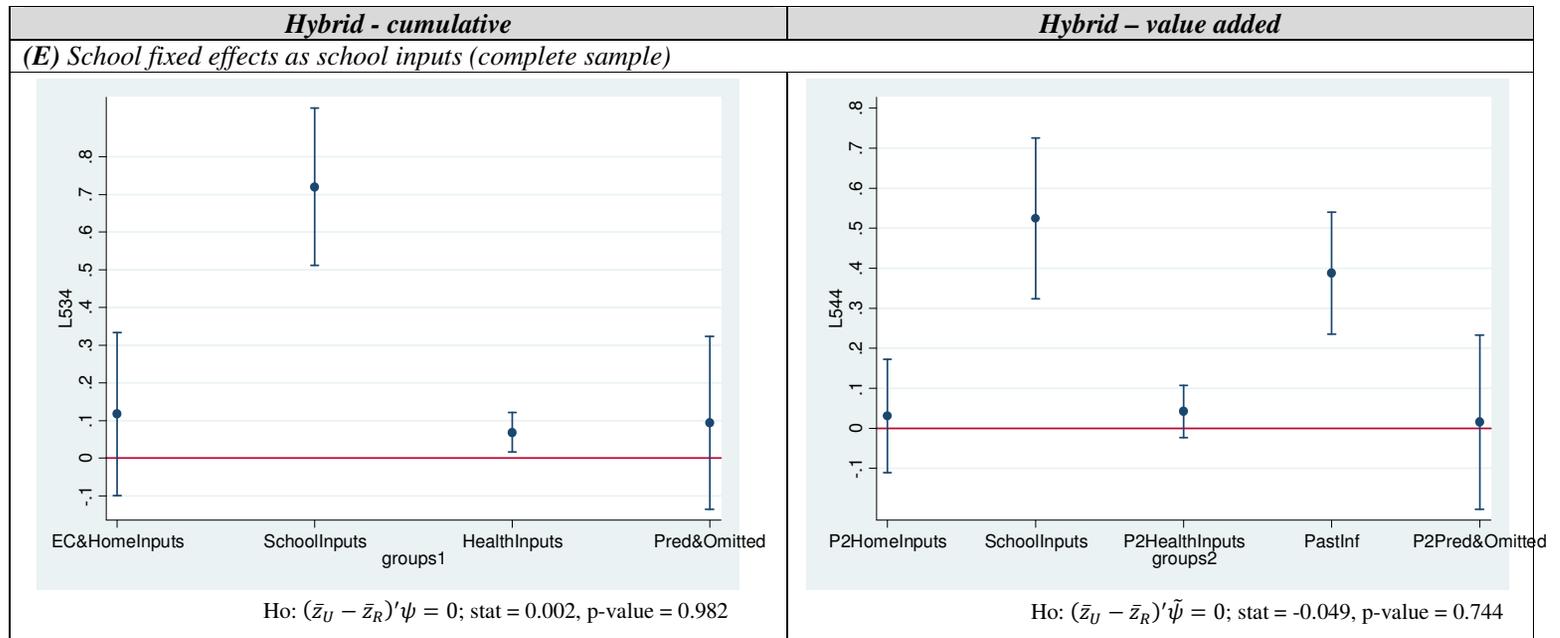
**Figure 1**  
**Normalized**  
**contributions to**  
**the urban/rural**  
**gap in cognitive**  
**skill at age 8**  
**(point estimates**  
**and 95%**  
**confidence**  
**intervals)**



**Figure 1**  
**Normalized**  
**contributions to**  
**the urban/rural**  
**gap in cognitive**  
**skill at age 8**  
**(point estimates**  
**and 95%**  
**confidence**  
**intervals)**



**Figure 1**  
**Normalized**  
**contributions to**  
**the urban/rural**  
**gap in cognitive**  
**skill at age 8**  
**(point estimates**  
**and 95%**  
**confidence**  
**intervals)**



**Notes:**

- “EC&HomeInputs” refers to early childhood and home inputs.
- “Pred&Omitted” refers to predetermined direct influences and omitted inputs.
- “P2HomeInputs” refers to period 2 home inputs.
- “P2HealthInputs” refers to period 2 health inputs.
- “PastInf” refers to past influences.
- “P2Pred&Omitted” refers to period 2 predetermined direct influences and period 2 omitted inputs.
- “P2FamilyInf” refers to period 2 family influences.

If the contribution of omitted inputs is captured by their demand equations when exogenous input determinants are included in the regression, 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 7 and of Figure 1 are consistent with this. In particular, the contribution of the category hosting omitted inputs in the hybrid-cumulative specification grows twice as large when school survey information is omitted (from 33% in the “full information” decomposition up to 65%). There is also strong evidence of the presence of omitted inputs as the contribution of exogenous input determinants in the hybrid cumulative model is now 51% (it was 22% in the “full information” decomposition) and highly significant ( $p < 0.00$ ).

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 school inputs provided by the school survey). Importantly, the contribution of exogenous input determinants now remains significant (30%;  $p < 0.00$ ) which implies we cannot accept the null  $(\bar{z}_U - \bar{z}_R)' \tilde{\psi} + \tilde{\theta} = 0$  (see Panel B in Figure 1). This result, which differs from the one obtained with the complete set of data, confirms there are still relevant period 2 influences omitted. This is consistent with the fact that we are intentionally omitting school inputs.

It is also worth mentioning that, consistent with the fact that exogenous input determinants and predetermined direct influences are controlling for the omitted school inputs, we only observe a small increase in the estimated contribution of the categories hosting home inputs. This increase, which could be regarded as a bias, remains well within standard errors in both specifications (compare panels A and B in Figure 1).

Results presented in Panel C of Table 7 and of Figure 1 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.

The fact that the variables included in the “predetermined direct influences and omitted inputs” category are capturing the contribution of omitted school inputs implies that assigning these variables to a category that does not correspond to the school environment will generate a bias. This is precisely what happens under the standard decomposition rule. All family, household and child characteristics are assigned to a single category (“family influences”) while the remaining urban-rural indicator is used to capture the “unexplained” part of the cognitive skill gap. The new category hosting “family influences” has a contribution between 20 and 25 percentage points larger than the one capturing early childhood and home inputs in the alternative decomposition strategy (compare panels C and D in Figure 1). 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.

It is worth noticing that there is also an important difference in the way one would interpret the portion of the gap that cannot be explained by the observed influences. To see this, let us refer to the decompositions based on the value added model in panels C and D of Figure 1. Under the standard decomposition rule (panel D) and in absence of structure regarding the source of the “unexplained” part of the gap, one would conclude that past achievement and the available family and school influences explain nearly 80% of the cognitive skill gap and

that most of this contribution has to do with influences that belong to the home environment. In addition, one could say that the remaining 20% of the gap remains unexplained and that this could be due to the omission of relevant skill determinants or to the fact that urban and rural children transform inputs into test scores differently (i.e. they exhibit different returns for a given set of inputs)<sup>28</sup>.

If we follow the alternative decomposition strategy (panel C), however, one would notice that a contribution similar to that of past influences (around 40%) can be attributed to period 2 predetermined direct influences and omitted inputs. This estimate is not only closer to the contribution of the school inputs we are intentionally omitting (which is around 35%) but its interpretation is much more informative as it explicitly acknowledges the presence of omitted inputs that belong to the second period.

Finally, panel E in Table 7 and Figure 1 present decomposition results using school fixed effects instead of the inputs contained in the school survey. A comparison against the “full information” decomposition (see Panel A) reveals a tendency to overstate the contribution of school inputs, especially in the absence of information on past achievement (i.e. in the hybrid-cumulative model). In fact, in the hybrid-cumulative model is quite clear that the school fixed effect has absorbed missing inputs from the early childhood period leaving the “predetermined direct influences and omitted inputs” category with an insignificant contribution, as opposed to what happened in the “full information” decomposition.

## 6. Concluding remarks

Linear decompositions based on the Oaxaca-Blinder technique are a fairly common way of attempting an estimate of the contribution of two or more categories of variables to the difference in cognitive achievement between children of different socioeconomic backgrounds. Two prominent examples of these categories are family and school influences.

In this paper, I have argued that performing such decompositions in absence of a framework postulating how cognitive skill is accumulated and how are its inputs determined, can be problematic in several ways. In particular, absence of this framework can lead one to overlook the difference between skill inputs and skill input determinants and to make arbitrary choices in terms of decomposition strategy and interpretation of its components. This, in turn, can lead to biases in the estimated contributions and to misleading policy implications.

This analysis has reviewed several studies using data from developing countries<sup>29</sup> and has found no consensus regarding the specific Oaxaca-Blinder strategy to use and how to

---

<sup>28</sup> Notice that for the standard rule I have employed an OB pooled\* decomposition. If we use an OB pooled decomposition (where reference coefficients are provided by a pooled regression that excludes the group indicator), we would obtain an even smaller estimate for the “unexplained” part of the gap (13%), which implies a larger risk of overstating the contribution of home influences.

<sup>29</sup> The analysis presented in Krieg and Storer (2006) is an interesting example from the developed world. The authors aim at determining how much of the difference between high and low-performing schools in the state of Washington can be attributed to characteristics that are beyond the control of school administrators. They used and OB twofold decomposition and found that a significant proportion of the gap can be attributed to characteristics of students and their families. Accordingly, they concluded arguing against penalizing poor performing schools. Interestingly, these authors acknowledge that student characteristics can pick-up the influence of omitted school inputs if sorting into different schools correlates with students’ socioeconomic

interpret the “unexplained” part of the gap. These studies also exhibit a tendency to group all observed family, household and child characteristics in a category different from the school environment (see, for example, McEwan and Marshall (2004) or Hernandez-Zavala et al. (2006)). I have argued this can lead to an overstatement of the importance of family and household influences because several of these characteristics can be controlling for omitted inputs that belong to the school environment. School fixed effects can pick-up the contribution of omitted inputs that belong to the home environment, especially in highly segregated school systems. There are, therefore, reasons to doubt the significant school contributions found in those studies that have used this empirical strategy (see, for example, McEwan (2004) or McEwan and Trowbridge (2007)).

Based on Glewwe and Miguel (2008), I developed a simple model explaining the skill formation process and how its inputs, including school characteristics, are determined by families’ choices. I then used these insights to illustrate, analytically, the potential biases that can be introduced by the rule of grouping all family, household and child characteristics into a single category. I also used the results of the model to justify the categories proposed for an alternative decomposition strategy that aims at being less susceptible to biases that those used so far in the literature.

This alternative strategy uses the coefficient estimates of a particular type of Oaxaca-Blinder decomposition, but arranges the contribution of individual variables considering that predetermined family, household and child characteristics belong to the demand functions of inputs. These variables are, therefore, grouped in a special category hosting omitted inputs that resembles the “unexplained” part of the gap in a more conventional Oaxaca-Blinder decomposition.

Finally, I illustrated empirically the main issues discussed in the analysis by decomposing the gap in cognitive achievement between urban and rural 8-year-old children in Peru. I relied on an unusually rich dataset containing comprehensive information on school characteristics and longitudinal information on skill that allows one to control for early childhood inputs and innate ability. This provided a fairly reliable estimate of the contribution of school inputs.

I then intentionally omitted information on school inputs and found that: (i) their contribution is picked-up by the predetermined family, household and child variables included, as predicted by the framework that understands these as arguments of their demand functions; (ii) assigning these exogenous input determinants into a single “family influences” category will lead to an overstatement of their contribution; (iii) the alternative decomposition strategy proposed here is less prone to this kind of bias and provides a more accurate interpretation of the “unexplained” part of the gap; and (iv) the use of school fixed effects leads to an overstatement of the contribution of school inputs, especially when information on early childhood influences is lacking.

If one seeks to perform a linear decomposition of an achievement gap, this analysis suggests one needs to exercise caution when classifying the potential covariates. In particular, it is always advisable to determine which of the available variables better reflect an input of skill and which better reflect an input determinant (i.e. an argument in the demand function of the

---

status. The authors called this the “Tiebout effect” based on Tiebout’s model where agents choose public goods by deciding where to live (Tiebout, 1956). In a similar fashion, the model with endogenous school quality proposed here also gives a role to families’ choices as a determinant of school inputs.

inputs of skill). Based on this, one can then assess which are the assumptions required for input determinants to control for omitted inputs that belong only to a certain category. For example, using predetermined family, household and child characteristics to control only for omitted home influences implies there are no omitted school inputs or that family decisions have not a significant role in determining the quality of the school environment. If these assumptions are not plausible considering the data in hand and the education system under analysis, one can rely on the decomposition strategy proposed here.

**Table 3.1**  
**Studies that have attempted a linear decomposition of the observed cognitive achievement gap between two groups of children**

	(A) Author(s), year and country	(B) Outcome and groups	(C) Empirical specification, decomposition strategy and components	(D) Categories proposed for observed influences	(E) Observed influences and data sources	(F) Results, remarks and policy implications
1	McEwan and Trowbridge (2007)  Guatemala	Test scores in language and maths  Indigenous vs. non-indigenous children in grades 3 and 6	Regressions on contemporaneous predictors.  OB twofold; reference coefficients from pooled regression including group indicator (pooled*)  Components: endowments ("explained"), group indicator ("unexplained")	C1: Family variables  C2: Quality of schools	C1: Parental education, presence of books, television viewing, child sex  C2: School fixed effects  PROENERE survey (2001)	> Explained component (language- mathematics) Grade 3: 71%-77% Grade 6: 55%-68%  > Categories (language-mathematics) Grade 3: C1= 6%-8%; C2 = 65%-69% Grade 6: C1 = 5%-3%; C2 = 50%-66%  Remarks: Between 50-69% of the gap is explained by the varying quality of schools that are attended by indigenous and non- indigenous children.
2	McEwan (2004)  Bolivia Chile	Test scores in language and maths  Indigenous vs. non-indigenous children in grades 3 and 6 (Bolivia) and 4 and 8 (Chile)	Regressions on contemporaneous predictors  OB twofold; reference coefficients from pooled regression including group indicator (pooled*)  Components: endowments ("explained"), group indicator ("unexplained")	C1: Family variables  C2: School variables  C3: Classroom variables	C1: Parental education, access to basic services (Bolivia), income (Chile), presence of books at home (Chile)  C2: School fixed effects  C3: Class fixed effects  Bolivia: SIMECAL (1997) Chile: SIMCE (1997-1999)	> Explained component Between 80% and 90%  > Categories C1 between 20% and 40% C2 and C3 between 50% and 70%  Remarks: Between 50% and 70% of the gaps are attributable to differences in the quality of schools and classrooms. The gap may be the result of an unequal distribution of school and classroom resources, but also of an unequal distribution of peer-group characteristics.

	(A) Author(s), year and country	(B) Outcome and groups	(C) Empirical specification, decomposition strategy and components	(D) Categories proposed for observed influences	(E) Observed influences and data sources	(F) Results, remarks and policy implications
3	Ramos et al. (2012)  Colombia	Test scores in mathematics, science, reading  Urban vs. rural students	Regressions on contemporaneous predictors  OB twofold; reference coefficients from rural group  Components: endowments ("explained"), coefficients ("unexplained")	C1: Individual characteristics  C2: Family characteristics  C3: School characteristics	C1: Child's age, child's sex  C2: Parental education, second generation migrant, home language, presence of books at home  C3: School is private/public, size, students per teacher, mean socioeconomic level of peer group ("peer effects")  PISA survey (2006 and 2009)	> Explained component 100% of gap ("unexplained" component has negative contribution)  > Categories C3 between 75% and 83% (mostly explained by the mean socioeconomic level of peer group)  Remarks: Most of the rural-urban school differential is related to family characteristics.  Policy implications: measures aimed at improving the general educational situation and conditions in the family.
4	McEwan and Marshall (2004)  Cuba Mexico	Test scores in language and maths  Mexican vs. Cuban children in grades 3 and 4	Regressions on contemporaneous predictors  OB twofold; reference coefficients from low- achieving country (Mexico) and high- achieving country (Cuba)  Components: endowments ("explained"), coefficients ("unexplained")	C1: Student and family variables  C2: Peer variables  C3: School and teacher variables	C1: Parental education, access to books, child's sex  C2: Class average of parental education and access to books, climate in class  C3: Grade, access to preschool, access to textbooks, teacher training and qualifications, access to materials, classroom conditions, school is urban/rural  UNESCO Assessment of Student Achievement (1997)	> Explained component Between 11% and 31%  > Categories C1 between 3% and 10% C2 between 14% and 23% C3 less than 3%  Remarks: No more than 30% of the difference can be explained by differences in family, peer and school characteristics. Peer group and family variables contribute the largest portion to the explained component. Observed school and teacher variables are inconsistently linked to achievement.

	(A) Author(s), year and country	(B) Outcome and groups	(C) Empirical specification, decomposition strategy and components	(D) Categories proposed for observed influences	(E) Observed influences and data sources	(F) Results, remarks and policy implications
5	Hernandez-Zavala et al. (2006)  Guatemala Mexico Peru	Test scores in language and maths  Indigenous vs. non-indigenous children between ages 8-10	Regressions on contemporaneous predictors  OB twofold; reference coefficients from non-indigenous group  Components: endowments ("explained"), coefficients ("unexplained")	C1: Family and child inputs  C2: School inputs	C1: Parental education, access to basic services (Mexico, Guatemala), household assets (Guatemala), books at home (Peru), child's sex, child's grade, child repeated a grade, child works, child attended preschool  C2: School is private/public; urban/rural (Peru, Guatemala), teacher experience, classroom condition (Peru and Mexico), access to textbooks (Guatemala), pupil-teacher ratio (Peru, Guatemala)  Peru: First Comparative International Study on Language, Mathematics and Associated Factors (1997) Guatemala: "Laboratorio Guatemala" (2002) Mexico: National Standards (2001)	> Explained component (language-mathematics) Guatemala: 41%-55% Mexico: 75%-68% Peru: 70%-66%  > Categories (language-mathematics) Guatemala: C1 = 23%-33%; C2 = 17%-23% Mexico: C1 = 67%-75%; C2 = 0% Peru: C1= 38%-41%; C2 25%-32%  Remarks: Family variables contribute more than school variables to the overall explained component.  Policy implications: effective bilingual education, compensatory education programs, choice of school and increased autonomy.

	(A) Author(s), year and country	(B) Outcome and groups	(C) Empirical specification, decomposition strategy and components	(D) Categories proposed for observed influences	(E) Observed influences and data sources	(F) Results, remarks and policy implications
6	Arteaga and Glewwe (2014)  Peru	Test scores in PPVT and maths  Indigenous vs. non-indigenous children at ages 8 and 5	Regressions on past and contemporaneous predictors. Separate models for children at 8 and 5 years of age  OB threefold; reference group is the indigenous group  Components: endowments, coefficients ("heterogeneous effects"), interaction	C1: Household and child characteristics  C2: Community characteristics	C1: Household expenditure, parental education, school expenditure, months in day care, months breastfeeding, prenatal visits, child did homework with parents, child played with parents, child's age, child's sex, child's nutritional status.  C2: Community fixed effects  Young Lives Study, rounds 2 (2006) and 3 (2009)	Most significant contributors: > By 5 years of age; PPVT Endowments of C2 = 67% > By 8 years of age; PPVT and maths Endowments of C1 > 80% (especially in parental education). * The contribution of interaction terms is not reported  Remarks: By age 8 the importance of community characteristics recedes and household and child characteristics play the major role.  Policy implication: Increase indigenous children's years of education as they will be household heads in the future.

	(A) Author(s), year and country	(B) Outcome and groups	(C) Empirical specification, decomposition strategy and components	(D) Categories proposed for observed influences	(E) Observed influences and data sources	(F) Results, remarks and policy implications
7	Barrera- Osorio et al. (2011)  Indonesia	Test scores in mathematics  2006 vs. 2003 results	Regressions on contemporaneous predictors  OB twofold; (a) reference coefficients from 2006 group; (b) reference coefficients from pooled regression including group indicator  Components: (a) endowments ("explained"), coefficients ("unexplained"); (b) endowments ("explained"), group indicator ("unexplained")	C1: Institutions / schools  C2: Student characteristics  C3: Family background	C1: School determines pedagogy, adequate supply of teachers, private-public, urban- rural, % repeating grade  C2: grade, child's age, child's sex  C3: Parental education, books present at home, access to computer, mother tongue  PISA survey (2003 and 2006)	> Unexplained component Between 63% and 92%  > Categories Most of the change in returns was in C2  Remarks: Most of the test score increase between 2003 and 2006 was due to changes in the returns to the characteristics rather than due to changes in the characteristics themselves.

	(A) Author(s), year and country	(B) Outcome and groups	(C) Empirical specification, decomposition strategy and components	(D) Categories proposed for observed influences	(E) Observed influences and data sources	(F) Results, remarks and policy implications
8	Zhang and Lee (2011) Korea Japan OECD countries	Test scores in mathematics, science, reading  Specific OECD countries vs. OECD average	Regressions on contemporaneous predictors  Resembles OB twofold; gaps are measured against OECD average using predictor means and coefficients from a regression containing all OECD countries  Components: endowments ("explained"), coefficients ("unexplained")	C1: Individual characteristics  C2: Study time and activities  C3: Family background  C4: School characteristics	C1: Grade, age, sex, immigrant, speaks foreign language, occupational aspiration  C2: time spent at school, homework is assigned  C3: parental education, parental occupational status, wealth, books present at home  C4: % of girls, % repeating grade, class size, student-teacher ratio, teacher qualifications, access to computers/internet, shortage of teachers by subject  PISA survey (2006)	> Explained component Varies considerably across countries; e.g. in maths: 0% (Germany), 51% (Korea), 100% (Japan)  > Categories (specific results reported for Korea and Japan) Korea: explained component attributable to differences in C4 (63%-83%) Japan: explained component attributable to differences in C2 and C4 (> 85%)  Remarks: Analysis on Korea and Japan illustrates how to identify factors that contribute most to the gap. If the observed gap is mainly due to the "unexplained" component, public policy needs to focus on broader and underlying economic, social and cultural differences.
9	Burger (2011)  Zambia	Test scores in reading  Urban vs. rural schools	Regressions on contemporaneous predictors  OB twofold; reference coefficients from rural schools  Components: endowments ("explained"), coefficients ("unexplained")	None	Predictors: household asset index, parental education, pupil-teacher ratio, proportion of students who spoke English  Southern Africa Consortium for Monitoring Educational Quality (SACMEQ II)	> Endowments ("resources"): 55% > Coefficients ("returns to resources"): 45%  Policy implication: Resource investment will not have the required impact unless the efficiency gap is also addressed.

	(A) Author(s), year and country	(B) Outcome and groups	(C) Empirical specification, decomposition strategy and components	(D) Categories proposed for observed influences	(E) Observed influences and data sources	(F) Results, remarks and policy implications
10	Beltran and Seinfeld (2012)  Peru	Test scores in reading and maths  Urban vs. rural students	Regressions on contemporaneous predictors.  OB threefold; rural students as reference group.  Components: endowments, returns, interaction.	None	Predictors: preschool attendance, district classified as poor, parental education, mother tongue, teacher qualifications, school infrastructure, access to internet at school, % repeating grade at school, private- public school, class time (minutes)  National Student Evaluation (2010) and School Census (2010)	> Endowments: 36% (reading); 22% (maths) > Coefficients/returns: 14.4% (reading); 4.8% (maths) > Interaction: 49.6% (reading); 73.2% (maths)  Policy implication: Adequate provision of resources has to be complemented with quality assurance and effective use.

## References

- Almond, D., & Currie, J. (2011). Human Capital Development before Age Five. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. 4b).
- 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.
- Barrera-Osorio, F., Garcia-Moreno, V., Patrinos, H. A., & Porta, E. (2011). *Using the Oaxaca-Blinder Decomposition Technique to Analyze Learning Outcomes Changes over Time: An Application to Indonesia's Results in PISA Mathematics*. Policy Research Working Paper 5584. The World Bank.
- 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.
- Burger, R. (2011). School effectiveness in Zambia: The origins of differences between rural and urban outcomes. *Development Southern Africa*, 28(2), 157-176.
- Castro, J. F., & Rolleston, C. (2015). *Explaining the Urban-Rural Gap in Cognitive Achievement in Peru: The role of Early Childhood and School Influences*. Young Lives Working Paper No. 139.
- 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. (2008). Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation. *The Journal of Human Resources*, 43(4), 738-782.
- 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.
- Cunha, F., Heckman, J., & Schenach, S. (2010). Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica*, 78(3), 883-931.
- Dunn, L., Padilla, E., Lugo, D., & Dunn, L. (1986). *Manual del Examinador para el Test de Vocabulario en Imágenes Peabody: Adaptación Hispanoamericana*. Minnesota: AGS.
- 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.
- 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.
- Helmers, C., & Patnam, M. (2011). The formation and evolution of childhood skill acquisition: Evidence from India. *Journal of Development Economics*, 95, 252-266.
- 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.
- Jones, F., & Kelly, J. (1984). Decomposing Differences Between Groups: A Cautionary Note of Measuring Discrimination. *Sociological Methods & Research*, 12, 323-343.
- Krieg, J., & Storer, P. (2006). How Much Do Students Matter? Applying the Oaxaca Decomposition to Explain Determinants of Adequate Yearly Progress. *Contemporary Economic Policy*, 24, 563-581.
- Lopez-Boo, F. (2009). *The Production Function of Cognitive Skills: Nutrition, Parental Inputs and Caste Test Gaps in India*. Young Lives Working Paper No. 55.
- 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.
- Neumark, D. (1988). Employers' Discriminatory Behavior and the Estimation of Wage Discrimination. *The Journal of Human Resources*, 23, 279-295.
- Oaxaca, R. (1973). Male–female wage differentials in urban labor markets. *International Economic Review*, 14(3), 693-709.
- 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.
- 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.
- Tiebout, C. (1956). A Pure Theory of Local Public Expenditures. *Journal of Political Economy*, 64, 416-424.
- 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.
- Winsborough, H., & Dickenson, P. (1971). *Components of negro-white income differences*. Proceedings of the American Statistical Association, Social Statistics Section: 6-8.
- Zhang, L., & Lee, K. A. (2011). Decomposing achievement gaps among OECD countries. *Asia Pacific Education Review*, 12, 463-174.

## Appendix 1: From the cumulative to the value added specification with age-dependent parameters

Consider the following linear version of the skill formation technology with age-dependent parameters.

$$A_{i2} = HI'_{i2}\gamma_1^2 + HI'_{i1}\gamma_2^2 + SI_{i2}\phi_1^2 + h_{i2}\varphi_1^2 + h_{i1}\varphi_2^2 + f'_i\lambda_{(2)} + \mu_{i0}\beta_{(2)} \quad (i)$$

Superscript 2 in all the parameters in (i) implies they are specific to period 2. This implies that the effect of a certain input on skill not only depends on the time elapsed since the application of the input but also on the specific period (or age) in which it was applied. Based on this, period 1 skill can be expressed as:

$$A_{i1} = HI'_{i1}\gamma_1^1 + h_{i1}\varphi_1^1 + f'_i\lambda^1 + \mu_{i0}\beta^1 \quad (ii)$$

We can subtract  $\rho A_{i1}$  from (i) to obtain:

$$A_{i2} - \rho A_{i1} = HI'_{i2}\gamma_1^2 + HI'_{i1}(\gamma_2^2 - \rho\gamma_1^1) + SI_{i2}\phi_1^2 + h_{i2}\varphi_1^2 + h_{i1}(\varphi_2^2 - \rho\varphi_1^1) + f'_i(\lambda_{(2)} - \rho\lambda^1) + \mu_{i0}(\beta_{(2)} - \rho\beta^1) \quad (iii)$$

If the effects of inputs decay at a rate  $\rho$ , the effects of period 1 inputs on period 2 skill equal  $\rho$  times the effect of period 1 inputs on period 1 skill. This means:  $\gamma_2^2 = \rho\gamma_1^1$ ,  $\varphi_2^2 = \rho\varphi_1^1$ ,  $\lambda_{(2)} = \lambda^2 + \rho\lambda^1$  and  $\beta_{(2)} = \beta^2 + \rho\beta^1$ . This, in turn, implies (iii) can be written as:

$$A_{i2} = \rho A_{i1} + HI'_{i2}\gamma_1^2 + SI_{i2}\phi_1^2 + h_{i2}\varphi_1^2 + f'_i\lambda^2 + \mu_{i0}\beta^2 \quad (iv)$$

The main text presents the case for age-invariant parameters with essentially the same results (see equations (24), (25) and (26)). The only difference is that parameters accompanying home, school and health inputs, as well as predetermined direct influences and innate ability, are no longer age or period-specific.

## Appendix 2: Descriptive statistics and urban/rural gaps

	Mean	SD	Urban	Rural	Diff.
Standardized raw PPVT score (round 3)	1.780	0.951	2.095	1.028	1.067*** (0.14)
Standardized raw PPVT score (round 2)	0.024	0.968	0.355	-0.766	1.121*** (0.13)
Real expenditure in child (learning materials and entertainment; round 2) <sup>(a)</sup>	0.274	0.364	0.342	0.112	0.23*** (0.049)
Mother had antenatal visits during pregnancy (yes = 1)	0.828	0.378	0.848	0.778	0.071* (0.038)
Maternal response to child cry was affectionate (yes = 1)	0.230	0.421	0.286	0.097	0.188*** (0.05)
Child attended formal preschool (yes = 1)	0.766	0.424	0.892	0.465	0.427*** (0.055)
Household has books and child is encouraged to read (yes = 1)	0.450	0.498	0.478	0.382	0.096 (0.06)
Household has a computer (yes = 1)	0.140	0.347	0.195	0.007	0.188*** (0.039)
Real expenditure in child (learning materials and entertainment; round 3) <sup>(a)</sup>	0.432	0.572	0.517	0.230	0.287*** (0.063)
Child receives help from parents when doing homework (yes = 1)	0.665	0.472	0.758	0.444	0.314*** (0.029)
Hours in a typical day the child spends playing	4.346	1.517	4.488	4.005	0.483** (0.218)
Hours in a typical day the child spends sleeping	9.931	0.978	9.988	9.796	0.192 (0.114)
Hours in a typical day the child spends studying	1.945	0.834	2.120	1.526	0.594*** (0.078)
Child is stunted (yes = 1; round 2)	0.316	0.465	0.207	0.576	-0.369*** (0.034)
Child is stunted (yes = 1; round 3)	0.189	0.392	0.120	0.354	-0.235*** (0.041)
Hours in a typical day the child spends at school	6.171	0.720	6.131	6.269	-0.138 (0.108)
Years of schooling (basic education)	2.374	0.544	2.429	2.243	0.186** (0.085)
CLIM: absence of problems in class (score 12-48)	32.736	6.567	33.760	30.298	3.462** (1.317)

	<b>Mean</b>	<b>SD</b>	<b>Urban</b>	<b>Rural</b>	<b>Diff.</b>
INF: school has basic services (yes = 1)	0.556	0.497	0.761	0.069	0.691*** (0.087)
ACT: average curricular coverage (% of topics covered in depth)	0.531	0.153	0.564	0.452	0.111*** (0.034)
ORG: teacher absenteeism (%)	0.025	0.111	0.012	0.057	-0.045 (0.031)
ORG: school has a psychologist (yes = 1)	0.179	0.383	0.248	0.014	0.234* (0.109)
ORG: school is “multigrade” (yes = 1)	0.187	0.390	0.073	0.458	-0.385*** (0.084)
TEA: more than 50% of teachers graduated from a university (yes = 1)	0.456	0.499	0.551	0.229	0.322*** (0.091)
Child’s caregiver has higher education (yes = 1)	0.179	0.383	0.245	0.021	0.224*** (0.037)
Caregiver’s age	34.569	6.843	34.172	35.514	-1.342 (0.804)
Child is male (yes = 1)	0.478	0.500	0.490	0.451	0.038 (0.048)
Child’s mother tongue is Spanish (yes = 1)	0.893	0.309	0.985	0.674	0.312** (0.104)
Child’s age in months	96.510	3.708	96.500	96.537	-0.037 (0.507)
Child lives in urban area (yes = 1)	0.704	0.457	1.000	0.000	1.000
Average household total income <sup>(a)</sup>	1.512	1.116	1.711	1.037	0.674*** (0.111)
Average household size	5.538	1.849	5.270	6.176	-0.906** (0.306)
Proportion of male siblings	0.495	0.333	0.490	0.506	-0.016 (0.026)
Child birth order	2.475	1.584	2.194	3.144	-0.949*** (0.198)
Caregiver aspiration for child is university education (yes=1)	0.655	0.476	0.743	0.444	0.299*** (0.065)

The number of observations is 487 for all variables.

(a) x 1,000 soles; 2006 prices in urban Lima.

Robust standard errors in parentheses.

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

### Appendix 3: Coefficient estimates for the eight specifications involved

**Table 3A**  
**Estimations with the reduced sample**  
**(including all variables and excluding school inputs)**

VARIABLES	“Full information” estimations		Excluding school inputs (reduced sample)	
	Hybrid-CU	Hybrid-VA	Hybrid-CU	Hybrid-VA
Real expenditure in child (learning materials and entertainment; round 2)	0.076 (0.090)	-0.022 (0.083)	0.180 (0.106)	0.038 (0.087)
Mother had antenatal visits during pregnancy (yes = 1)	0.132** (0.051)	0.121** (0.052)	0.136** (0.050)	0.119** (0.053)
Maternal response to child cry was affectionate (yes = 1)	0.077 (0.063)	0.025 (0.070)	0.101 (0.062)	0.035 (0.065)
Child attended formal preschool (yes = 1)	0.049 (0.094)	0.003 (0.080)	0.050 (0.121)	0.009 (0.113)
Household has books and child is encouraged to read (yes = 1)	0.210*** (0.057)	0.235*** (0.068)	0.207*** (0.052)	0.240*** (0.062)
Household has a computer (yes = 1)	0.087 (0.059)	0.064 (0.055)	0.133* (0.075)	0.099 (0.072)
Real expenditure in child (learning materials and entertainment; round 3)	0.062 (0.062)	0.036 (0.064)	0.077 (0.058)	0.041 (0.064)
Child receives help from parents when doing homework (yes = 1)	0.007 (0.093)	-0.041 (0.091)	0.027 (0.085)	-0.023 (0.085)
Hours in a typical day the child spends playing	-0.003 (0.021)	-0.012 (0.026)	0.032* (0.017)	0.015 (0.022)
Hours in a typical day the child spends sleeping	-0.032 (0.036)	-0.044 (0.038)	-0.011 (0.032)	-0.028 (0.035)
Hours in a typical day the child spends studying	0.045 (0.045)	0.024 (0.041)	0.085* (0.046)	0.047 (0.046)
Child is stunted (yes = 1; round 2)	-0.048 (0.085)	-0.002 (0.084)	-0.079 (0.101)	-0.025 (0.094)
Child is stunted (yes = 1; round 3)	-0.212 (0.123)	-0.184 (0.133)	-0.226 (0.134)	-0.198 (0.143)
Hours in a typical day the child spends at school	-0.075 (0.051)	-0.070 (0.053)	-0.042 (0.056)	-0.043 (0.054)
Years of schooling (basic education)	0.342*** (0.081)	0.253*** (0.075)	0.326*** (0.077)	0.227*** (0.060)
CLIM: absence of problems in class (score 12-48)	0.009** (0.004)	0.011** (0.004)	--	--
INF: school has basic services (yes = 1)	0.183** (0.069)	0.044 (0.064)	--	--
ACT: average curricular coverage (% of topics covered in depth)	0.521 (0.308)	0.388 (0.267)	--	--
ORG: teacher absenteeism (%)	-0.780* (0.380)	-0.761** (0.317)	--	--
ORG: school has a psychologist (yes = 1)	0.194** (0.079)	0.203** (0.087)	--	--
ORG: school is “multigrade” (yes = 1)	-0.292** (0.120)	-0.308*** (0.098)	--	--

VARIABLES	“Full information” estimations		Excluding school inputs (reduced sample)	
	Hybrid-CU	Hybrid-VA	Hybrid-CU	Hybrid-VA
TEA: more than 50% of teachers graduated from university (yes = 1)	0.090* (0.049)	0.012 (0.046)	--	--
Child’s caregiver has higher education (yes = 1)	0.135* (0.066)	0.017 (0.056)	0.169** (0.075)	0.025 (0.058)
Caregiver’s age	0.013** (0.005)	0.008 (0.005)	0.012** (0.005)	0.007 (0.005)
Child is male (yes = 1)	0.028 (0.096)	0.034 (0.083)	0.056 (0.094)	0.062 (0.078)
Child’s mother tongue is Spanish (yes = 1)	0.341** (0.144)	0.389** (0.152)	0.390** (0.145)	0.439** (0.170)
Child’s age in months	0.014 (0.011)	0.004 (0.009)	0.018 (0.011)	0.007 (0.010)
Child lives in urban area (yes = 1)	0.120 (0.113)	0.028 (0.096)	0.395*** (0.076)	0.202** (0.091)
Average household total income	0.018 (0.024)	0.010 (0.021)	0.033 (0.032)	0.021 (0.027)
Average household size	0.013 (0.026)	0.013 (0.026)	0.006 (0.028)	0.006 (0.028)
Proportion of male siblings	-0.032 (0.175)	-0.100 (0.173)	-0.137 (0.146)	-0.202 (0.138)
Child birth order	-0.097*** (0.026)	-0.085** (0.031)	-0.101*** (0.025)	-0.088*** (0.029)
Caregiver aspiration for child is university education (yes = 1)	0.058 (0.054)	0.026 (0.059)	0.112 (0.072)	0.065 (0.070)
Standardized raw PPVT score (round 2)		0.341*** (0.052)		0.363*** (0.046)
Constant	-1.362 (1.176)	0.465 (1.117)	-1.894 (1.152)	0.102 (1.172)
Observations	487	487	487	487
R-squared	0.557	0.608	0.510	0.573

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

**Table 3B**  
**Estimations with the complete sample excluding school inputs and with the reduced sample including school fixed effects**

VARIABLES	Excluding school inputs (complete sample)		School fixed effects (reduced sample)	
	Hybrid-CU	Hybrid-VA	Hybrid-CU	Hybrid-VA
Real expenditure in child (learning materials and entertainment; round 2)	0.122*** (0.036)	0.034 (0.028)	-0.022 (0.099)	-0.103 (0.096)
Mother had antenatal visits during pregnancy (yes = 1)	0.193*** (0.044)	0.139*** (0.039)	0.212*** (0.062)	0.205** (0.071)
Maternal response to child cry was affectionate (yes = 1)	0.042 (0.043)	0.012 (0.046)	0.038 (0.062)	0.012 (0.056)
Child attended formal preschool (yes = 1)	0.059 (0.058)	0.016 (0.044)	0.070 (0.169)	0.043 (0.147)
Household has books and child is encouraged to read (yes = 1)	0.183*** (0.048)	0.194*** (0.048)	0.283** (0.105)	0.272** (0.110)
Household has a computer (yes = 1)	0.130*** (0.040)	0.082** (0.034)	0.074 (0.090)	0.061 (0.083)
Real expenditure in child (learning materials and entertainment; round 3)	0.059 (0.044)	0.029 (0.032)	0.145 (0.098)	0.098 (0.108)
Child receives help from parents when doing homework (yes = 1)	0.076 (0.045)	0.061 (0.044)	0.020 (0.149)	-0.023 (0.136)
Hours in a typical day the child spends playing	0.037* (0.021)	0.021 (0.018)	-0.037 (0.032)	-0.051 (0.031)
Hours in a typical day the child spends sleeping	-0.019 (0.024)	-0.036 (0.024)	-0.038 (0.032)	-0.054 (0.038)
Hours in a typical day the child spends studying	0.065*** (0.022)	0.017 (0.018)	0.025 (0.069)	0.015 (0.063)
Child is stunted (yes = 1; round 2)	-0.098 (0.066)	-0.024 (0.050)	-0.054 (0.089)	-0.022 (0.097)
Child is stunted (yes = 1; round 3)	-0.125** (0.052)	-0.111** (0.053)	-0.227* (0.122)	-0.190 (0.152)
Hours in a typical day the child spends at school	0.020 (0.037)	0.005 (0.034)	-0.048 (0.060)	-0.057 (0.058)
Years of schooling (basic education)	0.254*** (0.044)	0.134*** (0.034)	0.309** (0.060)	0.286** (0.060)
CLIM: absence of problems in class (score 12-48)	--	--	--	--
INF: school has basic services (yes = 1)	--	--	--	--
ACT: average curricular coverage (% of topics covered in depth)	--	--	--	--
ORG: teacher absenteeism (%)	--	--	--	--
ORG: school has a psychologist (yes = 1)	--	--	--	--
ORG: school is "multigrade" (yes = 1)	--	--	--	--
TEA: more than 50% of teachers graduated from university (yes = 1)	--	--	--	--

VARIABLES	Excluding school inputs (complete sample)		School fixed effects (reduced sample)	
	Hybrid-CU	Hybrid-VA	Hybrid-CU	Hybrid-VA
Child's caregiver has higher education (yes = 1)	0.194*** (0.058)	0.054 (0.048)	0.071 (0.101)	-0.059 (0.088)
Caregiver's age	0.009** (0.004)	0.006 (0.004)	0.018* (0.010)	0.014 (0.010)
Child is male (yes = 1)	0.068 (0.059)	0.060 (0.038)	0.021 (0.114)	0.056 (0.102)
Child's mother tongue is Spanish (yes = 1)	0.278** (0.105)	0.367*** (0.100)	0.343 (0.264)	0.316 (0.268)
Child's age in months	0.012 (0.007)	0.008 (0.006)	0.016 (0.009)	-0.002 (0.009)
Child lives in urban area (yes = 1)	0.453*** (0.085)	0.246*** (0.083)	-0.101 (0.152)	-0.131 (0.150)
Average household total income	0.048** (0.018)	0.027* (0.014)	0.001 (0.040)	-0.000 (0.032)
Average household size	0.009 (0.014)	0.008 (0.013)	0.039 (0.040)	0.042 (0.042)
Proportion of male siblings	-0.012 (0.069)	-0.029 (0.052)	-0.149 (0.219)	-0.218 (0.225)
Child birth order	-0.076*** (0.020)	-0.051** (0.021)	-0.105** (0.042)	-0.097* (0.048)
Caregiver aspiration for child is university education (yes = 1)	0.216*** (0.034)	0.147*** (0.033)	0.122 (0.081)	0.073 (0.083)
Standardized raw PPVT score (round 2)		0.399*** (0.030)		0.352*** (0.064)
Constant	-1.667* (0.807)	-0.249 (0.813)	-0.614 (0.697)	1.851* (0.895)
School fixed effects	No	No	Yes	Yes
Observations	1,561	1,561	487	487
R-squared	0.473	0.557	0.698	0.730

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1