139

Working Paper

Young Lives
An International Study of Childhood Poverty

Explaining the Urban–Rural Gap in Cognitive Achievement in Peru:

The Role of Early Childhood Environments and School Influences

Juan F. Castro and Caine Rolleston

www.younglives.org.uk

AUGUST 2015

139

Working Paper

Young Lives
An International Study of Childhood Poverty

Explaining the Urban–Rural Gap in Cognitive Achievement in Peru:

The Role of Early Childhood Environments and School Influences

Juan F. Castro and Caine Rolleston

www.younglives.org.uk

AUGUST 2015

Explaining the Urban–Rural Gap in Cognitive Achievement in Peru: The Role of Early Childhood Environments and School Influences

Juan F. Castro and Caine Rolleston

ISBN 978-1-909403-53-6

A catalogue record for this publication is available from the British Library.

Printed on FSC-certified paper from traceable and sustainable sources.

Funded by

UKaid
from the British people

Ministry of Foreign Affairs of the Netherlands

Irish Aid
An Roinn Gnóthaí Eachtracha agus Trádála
Department of Foreign Affairs and Trade

Young Lives, Oxford Department of International Development (ODID), University of Oxford,
Queen Elizabeth House, 3 Mansfield Road, Oxford OX1 3TB, UK

Tel: +44 (0)1865 281751 • E-mail: younglives@younglives.org.uk

Contents

Abstract

In Peru, students attending rural schools demonstrate extremely poor learning outcomes and obtain results significantly below those of students in urban schools. Because the process of cognitive skill formation is cumulative, differences in initial endowments, early environments and influences occurring later at home and at school can all play a role in shaping these gaps. This analysis aims at measuring the contribution of school and early childhood influences to the difference in cognitive development observed, at the age of 8, between urban and rural children in Peru. Previous decomposition exercises using Peruvian data on the indigenous–non-indigenous achievement gap, report results that favour the role of household characteristics over that of schools or community-level variables. This analysis contributes new evidence based on an unusually rich dataset and provided by a decomposition strategy less prone to biases than those used so far in the literature. Results indicate that between 35 and 40 per cent of the gap in cognitive skill between urban and rural 8-year-old children is related to differences in school inputs (years of schooling, school and teacher characteristics) received between the ages of 6 and 8. This contribution is similar to that of the learning and care environment to which the child was exposed up until the age of 5. The characteristics of rural schools have a direct connection with policy action because nearly all the supply of educational services in rural areas is public. Thus, efforts devoted to ensuring the characteristics of rural schools and teachers become more equal with those in urban areas should allow a significant reduction in the cognitive skill gap between urban and rural children by the time they reach Grade 3.

The Authors

Juan F. Castro is Associate Professor in the Economics Department at the Universidad del Pacifico in Lima. He holds an MSc degree in Economics from the London School of Economics and Political Science and is currently finishing his research to obtain a DPhil in International Development from the University of Oxford. His current work focuses on early childhood development, education, and applied micro-econometrics.

Caine Rolleston is a Senior Lecturer at the Institute of Education at University College, London and Senior Education Associate at Young Lives. His research interests focus on educational access, learning metrics, educational effectiveness and the economic benefits of education.

Acknowledgements

The authors would like to thank Stefan Dercon, Paul Glewwe, Douglas Gollin, Imane Chaara, Andreas Georgiadis and participants at the Queen Elizabeth House Development Economics Seminars for their comments. As usual, any remaining errors are our own.

About Young Lives

Young Lives is an international study of childhood poverty, following the lives of 12,000 children in 4 countries (Ethiopia, India, Peru and Vietnam) over 15 years. **www.younglives.org.uk**

Young Lives is funded from 2001 to 2017 by UK aid from the Department for International Development (DFID), co-funded by the Netherlands Ministry of Foreign Affairs from 2010 to 2014, and by Irish Aid from 2014 to 2015.

The views expressed are those of the author(s). They are not necessarily those of, or endorsed by, Young Lives, the University of Oxford, DFID or other funders.

1. Introduction and motivation

In Peru, students attending rural schools demonstrate extremely poor learning outcomes and obtain results significantly below those of students in urban schools. Among rural second-grade students, only 7 and 4 per cent demonstrate 'adequate' reading and mathematical skills, respectively, as compared to 38 and 15 per cent in urban areas (MINEDU 2013). Poverty incidence is also appreciably higher in rural areas, where 56 per cent of the population lives below the national poverty line, versus only 18 per cent in urban areas (INEI 2012).

Learning outcomes measured through reading comprehension and mathematics tests can be understood as a reflection of a broader process of cognitive skill formation. This process is cumulative and, therefore, achievement gaps such as the one documented above must be understood as the result of the effects of all relevant influences on cognitive development until the time at which outcomes are measured. In principle, differences in initial endowments, early environments and influences exerted later both at home and at school can all play a role in shaping these gaps.

Considerable debate surrounds the measurement of the contributions of home background and school quality to children's cognitive development and learning outcomes. Pioneering results were presented in the Coleman Report (Coleman et al. 1966) and the Plowden Report (Peaker 1971), which used data from the United States and England, respectively. These results suggested that influences occurring prior to school were more significant than school resources when explaining differences in student outcomes. Two widely cited articles, published several years after, employed data from developed and developing countries and found that the relationship between school resources and learning outcomes was negatively correlated with national per capita income (Heyneman and Loxley 1982, 1983). These authors found that school characteristics mattered more than family background in low-income countries. More recent studies have failed to confirm the negative association between national income and the proportion of variance in learning outcomes explained by school characteristics, and have questioned the extent to which, when fully accounting for backgrounds, school resources are effectively translated into improved cognitive outcomes (Baker et al. 2002; Hanushek and Luque 2003).

Several empirical efforts have tried to address this debate by decomposing learning-outcome gaps between children of dissimilar backgrounds into different groups or categories of variables. These studies have distinguished, primarily, between 'family' and 'school' influences and, in Latin America, have focused on learning gaps between indigenous and non-indigenous children (Arteaga and Glewwe 2014; Hernandez-Zavala et al. 2006; McEwan 2004; McEwan and Trowbridge 2007).

Results from the four studies cited above are mixed, which fails to clarify the role played by schools in shaping existing learning-outcome gaps and how it compares to the role played by influences exerted earlier in the lives of children. For example, results can vary between 17 per cent (Hernandez-Zavala et al. 2006) and 70 per cent (McEwan and Trowbridge 2007) when it comes to measures of the contribution of schools to the difference in learning outcomes between indigenous and non-indigenous children in Guatemala.

Two of the four cited studies employ data from Peru. In Hernandez-Zavala et al. (2006), the authors considered test scores in mathematics and language obtained by children between the ages of 8 and 10 in Guatemala, Mexico and Peru. They measured the contribution of 'family and child inputs' (such as parental education, household assets, presence of books at home, and pre-school attendance) and 'school inputs' (such as teacher experience, access to textbooks, and pupil–teacher ratios) to the indigenous–non-indigenous achievement gap. They found that, in general, family variables contributed more than school variables (23–33 per cent vs. 17–23 per cent in Guatemala; 67–75 per cent vs. 0 per cent in Mexico; 38–41 per cent vs. 25–32 per cent in Peru).

In a more recent analysis, Arteaga and Glewwe (2014) used test scores from the Peabody Picture Vocabulary Test (PPVT) and a mathematics test taken by the same group of children at the ages of 5 and 8. They measured the contribution of 'household and child characteristics' (such as household expenditure, parental education, if parents helped with homework and children's nutritional status) and 'community characteristics' (captured through community fixed effects) to the indigenous–non-indigenous achievement gap present at both ages. They found that, by the time children were enrolled in school (at the age of 8), household and child characteristics accounted for a significant portion of the achievement gap (around 80 per cent) while community-level variables appeared to play only a minor role.

Determining which influences and environments play a larger role in shaping achievement gaps can be informative for policy action aimed at overcoming developmental setbacks affecting children from disadvantaged backgrounds. Strong correlations between school characteristics and family backgrounds are a feature of many developing countries and Peru is no exception (Beltran and Seinfeld 2012; Cueto et al. 2013; Cueto et al. 2014a, 2014b). This potential for mutually reinforcing forms of disadvantage means such enquiry is especially important from an equity perspective, in relation to the role of schooling in mitigating the effects of disadvantage caused by home background.

In this regard, the few results available for Peru and reported above tend to favour the role of household characteristics over that of schools or community-level variables. These results seem supported by the fact that gaps in cognitive achievement between children from disadvantaged and more affluent backgrounds emerge before they enter school and persist over time (Schady et al. 2014). Accordingly, policy recommendations aimed at reducing achievement gaps end up focused on variables such as parental education (see, for example, Arteaga and Glewwe 2014). Before concluding that school influences play a subsidiary role when compared to early parental investments and home environments, however, we need to acknowledge some empirical challenges that have not yet been fully addressed by the literature.

In fact, strong correlations between school characteristics and family backgrounds also render the effort of comparing the relative importance of early childhood environments and school influences more challenging as the presence of confounders is more likely. In particular, one faces two types of bias when attempting an exercise such as the ones reviewed above. The first is related to the presence of bias in the estimates of individual effects of the determinants of skill. The second can emerge when assigning these determinants to different categories or groups of variables and is caused by the use of rules that end up assigning part of the contribution of one category to another.

Both types of bias have their origin in the presence of unobserved influences. Due to the cumulative nature of the cognitive skill formation process, one needs a particularly rich database (accounting for past and contemporaneous influences) to be able to minimise the risk of omitted variable bias when estimating individual effects. In this regard, three of the four studies cited above relied only on cross-sectional datasets that lacked information on past influences affecting the acquisition of skill.

In addition, the objective of estimating the contribution of different categories of variables to a certain achievement gap and the use of variables that control for omitted influences (such as family income or parental education) pose another challenge. As documented above, is common practice to assign these controls to the category containing 'family' influences as these variables typically comprise household characteristics. This entails the risk of introducing another type of bias by overstating the contribution of 'family' or 'household' influences. This is because these variables can also control for omitted inputs that belong to the school environment. As will be discussed later in this paper, it is useful to distinguish between skill inputs (i.e. variables that have a direct effect on skill such as educational materials provided at home) and skill input determinants (such as family income) when devising a decomposition strategy, in order to minimise this second type of bias.

From the discussion above is clear that there is still an unresolved question regarding the relative importance of schools and early childhood influences for unequal developmental outcomes among children living in the developing world. This study aims to advance the literature on skill formation and child development by addressing this issue in the Peruvian context using: (i) a comprehensive dataset that includes longitudinal information on cognitive skills and a wide range of child and household characteristics, combined with detailed information at the school level; and (ii) a decomposition strategy which acknowledges the difference and relations between skill inputs and skill input determinants. The first feature allows one to consider results from several empirical specifications and reduce the risk of obtaining a biased estimate of the individual effect of observed influences. The second feature makes the analysis less prone to the second type of bias described above.

The rest of the paper is organised as follows. Section 2 summarises key issues of access, learning outcomes and inequalities in the Peruvian system of basic education. Section 3 presents the 'skill formation technology' and further discusses the two types of bias that can affect an empirical exercise aimed at decomposing an achievement gap. Section 4 introduces our decomposition strategy. The results are presented and discussed in Section 5, and Section 6 closes with concluding remarks and policy implications.

2. Educational inequality in Peru

Peru has achieved high levels of enrolment in education, especially at primary level, where the net enrolment rate stood at 94 per cent in 2011 (World Bank 2014). Both primary and secondary education, which cover the ages of 6 to 16, are compulsory and free of tuition fees in public schools, following the 2003 General Education Law. At least one year of pre-schooling (which is compulsory), provided by the state and by private providers, is accessed by a vast majority of children and coverage is being extended to children from the age of 3. The gross enrolment ratio for pre-schooling overall was 78 per cent in 2012 (World Bank 2014). Despite high enrolment from a comparatively young age, however, available evidence suggests that learning levels in Peru are generally low, as well as being highly unequal.

In international testing exercises, Peru's ranking is among the lowest among participating countries in Latin America and is below the levels found in comparable middle-income countries (World Bank 2007). In the recent Programme for International Student Assessment

(PISA) 2012 exercise, Peru ranked last in mathematics, reading and science out of 65 countries (OECD 2013).

Economic conditions alone do not explain poor performance, drawing attention to issues of school quality, especially in the public school system. Within this system, the World Bank highlights issues related to weaknesses of management and accountability, pedagogical development with regard to teaching in disadvantaged areas, and issues of linguistic and cultural discrimination (World Bank 2007).

In addition to general issues of school quality, part of the explanation for low overall levels of performance lies in the 'long tail' of very poorly performing pupils and schools found in a highly unequal system. In comparative terms, Peru's test scores are among the most unequally distributed among all countries for which data are available, with levels of inequality being similar to those in South Africa, which endured decades of enforced discrimination, although there is some evidence of improvement over time (Crouch et al. 2009).

Inequalities in learning outcomes are accompanied by an unequal distribution of educational inputs. While public expenditure on basic education is not especially regressive,[1] the tendency of wealthier parents to choose private schools and highly unequal levels of private spending on education contribute to a regressive distribution of infrastructural and supply inputs across the sector. Recent studies show significant gaps in 'opportunities to learn' (OTL)[2] between more and less advantaged pupils in Peru, along axes of language and indigenousness, socioeconomic status and urban/rural residence (Cueto et al. 2014a; Cueto et al. 2014b).

In Table 1 we provide additional evidence regarding inequalities in educational inputs at the school level, using the school survey data employed in the decomposition exercises that follow. This school survey was conducted in 2011 as part of the Young Lives study in Peru.[3] It contains rich information on school and teacher characteristics from a random sample of schools attended, at the age of 10, by the Younger Cohort of children followed by Young Lives.

As discussed above, disparities are not especially wide in terms of basic access. In fact, average years of schooling at the age of 8 are between 2.2 and 2.6 years if we compare the lowest and highest wealth quartiles.[4] There is, however, a strong correlation between children's socioeconomic status (SES) and the quality of their schooling. In particular, access to basic services and educational infrastructure, teacher qualifications, the use of learning material and curriculum coverage all exhibit positive and significant SES gradients. Increases in school quality as we move up in the wealth distribution are linked to a shift from rural to urban settlement and to higher enrolment rates in private schools. Wealthier families can

[1] As noted in Crouch et al. (2009), this is not a result of policy but of private choice: wealthier families self-select out of the public system.

[2] OTL typically reflects curriculum content and coverage. In Cueto et al. (2014a) authors employed four variables to account for OTL: curriculum coverage, hours of mathematics at school per year, quality of teachers' feedback and the level of cognitive demand of exercises.

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

[4] Based on a wealth index comprising dwelling characteristics, access to basic services and durable goods consumption.

afford to pay more for privately provided educational services and also have access to better public schools in urban areas.

Importantly, children's 'school readiness' (measured by their cognitive achievement prior to commencing primary education) also exhibits a positive and highly significant SES gradient. As discussed in the preceding section, this combination of mutually reinforcing forms of disadvantage renders the effort of comparing the importance of early childhood and school influences both relevant in terms of policy and challenging in terms of empirical strategy.

Table 1. *Children's school readiness, years of schooling, and school and teacher characteristics by student's wealth quartile (Q)*

	Q1	Q2	Q3	Q4
Household is urban (%)	0.288	0.509	0.944	0.992
		0.221	0.656***	0.704***
Standardised raw PPVT score (age 5)	−0.776	-0.478	0.322	0.87
		0.298***	1.099***	1.646***
Years of schooling (age 8)	2.208	2.228	2.458	2.553
		0.02	0.25**	0.345***
School infrastructure and organisation (%)				
School is private	0.00	0.044	0.146	0.268
		0.044	0.146**	0.268***
School has basic servicesa	0.120	0.281	0.833	0.894
		0.161*	0.713***	0.774***
School has a library	0.256	0.456	0.66	0.659
		0.200***	0.404***	0.403***
School has a sports ground or playground	0.552	0.667	0.896	0.935
		0.115	0.344***	0.383***
School is multigradeb	0.480	0.211	0.069	0.057
		−0.269**	-0.411***	−0.423***
Teacher characteristics				
Teacher experience (years)	16.165	18.414	20.548	19.118
		2.249	4.384**	2.953
Teachers have a university degree (%)	0.349	0.415	0.525	0.703
		0.066	0.176**	0.354***
Material and curriculum (%)				
Teachers use textbooks and workbooks	0.189	0.234	0.34	0.359
		0.045	0.151***	0.169***
Teachers use complementary materials	0.177	0.26	0.313	0.31
		0.083***	0.136***	0.133***
Curriculum coverage (in depth)	0.456	0.504	0.55	0.597
		0.048	0.095**	0.141***
Number of obs.	125	114	144	123

Second row for each variable indicates the difference with respect to Q1

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

a Basic services comprise water (from a public network or pipe), sanitation (public network connection or a treated cesspool), electricity and telephone connection.

b 'Multigrade' means that children from different grades receive classes at the same time, in the same room, and from the same teacher.

3. Determinants of cognitive skill and empirical challenges for a gap decomposition

3.1. The production function of skill

The main objective of this analysis is to estimate the contribution of direct influences occurring at school and of direct influences exerted during early childhood, to the difference in cognitive skill between urban and rural school-age children in Peru.

Consistent with this objective, let us divide the relevant phase of child development into two time periods. The first begins when the child is born and finishes at the age of 5, that is, when the child is ready to start the basic education cycle and enrol in primary school. The second period corresponds to the time when the child remains within primary school age, which is usually between the ages of 6 and 11.

The production function of cognitive skill provides a good starting point to define its determinants. We assume that 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:

$$A_{i2} = A_2\left(HI_{i2}, HI_{i1}, SI_{i2}^{j}, SY_{i2}^{j}, h_{i2}, h_{i1}, f_i, \mu_{i0}\right) \qquad (1)$$

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}^{j} are educational inputs provided at the school where the child is enrolled (school j) during Period 2; SY_{i2}^{j} are years of schooling attained at school j during Period 2[5]; h_{it} indicates the child's health status during period t; f_i captures predetermined direct influences; and μ_{i0} is the child's 'innate ability'.

Importantly, equation (1) denotes a structural relationship between skill and those variables that have a *direct* effect on it. These variables will reflect the environment surrounding the child (characterising 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, we further classify these direct influences as inputs (if they are determined by families' choices during the period under analysis) or as predetermined (if they are outside the current choice set of families).

According to this formulation, differences in terms of cognitive achievement between children living in rural and urban areas should be related to differences between these geographical domains in terms of the direct influences indicated in (1).[6] Complete and accurate data on skills and on all these direct influences would allow one to: (i) estimate the structural relationship (i.e. estimate the individual effect of each direct influence); and

[5] We assume children do not switch schools during Period 2.

[6] Notice that we are assuming that the functional form and parameter values of the production function of skill are the same for urban and rural children. One way of interpreting this is that we are assuming there are no biological differences between children living in these two geographical domains.

(ii) estimate the contribution to the urban–rural gap of different groups or categories of influences.

To see this, let us define the urban–rural gap in cognitive skill as the difference in average skill between urban and rural children by the end of Period 2 $\left(E(A_{i2}|U) - E(A_{i2}|R)\right)$ and consider a linear version of (1) such as:

$$A_{i2} = HI'_{i2}\gamma_1 + HI'_{i1}\gamma_2 + SI^j_{i2}\phi_1 + SY^j_{i2}\alpha_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f_i{}'\lambda_{(2)} + \mu_{i0}\beta_{(2)} \qquad (2)$$

Accordingly, the contribution of school influences (CS_2) and of inputs provided during early childhood (CE_2) to the urban–rural gap can be defined as follows:[7]

$$CS_2 = \left(E\left(SI^j_{i2}|U\right) - E\left(SI^j_{i2}|R\right)\right)'\phi_1 + \left(E\left(SY^j_{i2}|U\right) - E\left(SY^j_{i2}|R\right)\right)\alpha_1 \qquad (3)$$

$$CE_2 = \left(E(HI_{i1}|U) - E(HI_{i1}|R)\right)'\gamma_2 + \left(E(h_{i1}|U) - E(h_{i1}|R)\right)'\varphi_2 \qquad (4)$$

Sample means of $SI^j_{i2}, SY^j_{i2}, HI_{i1}$ and h_{i1} from the urban and rural domains, together with estimates of parameters $\phi_1, \alpha_1, \gamma_2$ and φ_2, could be employed to calculate the contributions expressed in (3) and (4), and provide an answer to the research question proposed for this analysis.

Unfortunately, complete and accurate information on all the variables considered in (1) is seldom available. This is not only because the complete set of relevant direct influences is difficult to define, but also because several of these influences are difficult to measure (e.g. teachers' pedagogical practices) while others remain unobserved (e.g. child's innate ability).

Omission of relevant direct influences can lead to biased estimates of the contribution of different categories of variables in two ways. These two types of bias are related to the two basic ingredients of an empirical exercise aimed at decomposing an achievement gap:
(i) knowledge of the individual effects of the direct influences of skill; and (ii) different variable categories and a rule to assign the contribution of particular influences to them.

Accordingly, the first type of bias is caused by the presence of biased estimates of the individual effects of one or more variables. The second type is related to the use of rules (explicit or implicit) that assign the contribution of variables that belong to one category to another. In terms of the categories proposed for this study, this would mean that at least part of the contribution of the school influences category is being captured by the group referred to the early childhood environment, or vice versa. In what follows, we further discuss these two types of bias and summarise our approach to minimise their effects.

3.2. Estimating individual effects

Correlation between included direct influences and those omitted will lead to biased estimates of individual effects. This correlation is likely because skill inputs are related through the decision-making process of families. The empirical strategy to be used to overcome this problem has to take into account the nature of the parameters of interest.

The decomposition exercise proposed for this study requires parameter estimates for all relevant direct influences indicated in (2); that is, we require an estimate of the parameters of the production function of skill. This detail is not trivial as it implies that our identification strategy for the individual effects of direct influences has to be based on employing a rich set

[7] Here we refer to the contribution of *inputs* provided during early childhood because we are not considering the effect of predetermined direct influences and innate ability.

of controls and imposing some assumptions on the production function of skill and the behaviour of its determinants.

Experimental designs and quasi-experimental techniques are not only impractical but also unsuited to estimating the parameters of interest. Lack of practicality stems from the difficulty of randomly assigning children to receive several different treatments (one for each input under analysis plus a control group) or, if we aim at using a 'natural experiment', it stems from the difficulty of finding enough sources of exogenous variation (i.e. enough valid instruments).

Equally, or more, important is the fact that these techniques are particularly well suited to estimating the *total* effect of a particular input and not to estimating its production function parameter or *direct* effect. Parameters involved in (2) indicate a direct effect because they capture the result of shifting a particular influence holding the rest constant. Random assignment to a treatment group or a valid instrumental variable, however, will provide an estimate of this direct effect plus an indirect one which occurs through changes in other inputs (recall that inputs are choice variables and can therefore change in response to a shock to some skill determinant).[8]

In Todd and Wolpin (2003) and Todd and Wolpin (2007), the authors have already discussed the assumptions required to obtain consistent estimates of production function parameters under different empirical specifications. Here we briefly review the assumptions related to two particular specifications (the hybrid model and the value-added model) as they will play an important role in our empirical strategy.

3.1.1. The hybrid model

Let us return to the linear production function specified in (2) and drop the school identifier to simplify notation. We also assume that skill is measured with error through the scores obtained in some cognitive test: $T_{i2} = A_{i2} + \varepsilon_{i2}$. Thus, we obtain:

$$T_{i2} = HI'_{i2}\gamma_1 + HI'_{i1}\gamma_2 + SI'_{i2}\phi_1 + SY_{i2}\alpha_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f_i'\lambda_{(2)} + \mu_{i0}\beta_{(2)} + \varepsilon_{i2} \quad (5)$$

As already noted, consistent estimation of the parameters involved in the expression above is problematic because we seldom observe all the relevant direct influences. A common practice is to try to circumvent this problem by substituting missing inputs with their corresponding demand functions, which will typically contain variables reflecting family resources, prices, predetermined direct influences of skill, exogenous environmental characteristics, child's skill and health endowments, and parental preferences.[9] This gives rise to what the literature has described as a 'hybrid' function (Rosenzweig and Schultz 1983; Todd and Wolpin 2007).

To illustrate this, assume there are only two Period 1 educational inputs: one that can be observed (HI_{i1}^{O}) and one that cannot (HI_{i1}^{U}). Additionally, assume that the demand for the unobserved input can be expressed as a linear function of the predetermined direct influences of skill (f_i) and a vector of exogenous input determinants (z_i), up to an error term

[8] This means that experimental and quasi-experimental techniques usually recover the parameters of a conditional demand function and not the parameters of a production function. See Glewwe and Miguel (2008) for a discussion of the differences between these two types of functions. In Todd and Wolpin (2003) the authors distinguish between 'policy effects' and 'technology parameters' and provide some examples.

[9] See Glewwe and Miguel (2008) for a good exposition regarding input demand functions.

(v_i): $HI_{i1}^U = z_i'\delta + f_i'\kappa + v_i$. This enables us to replace the unobserved input for its demand function to obtain:

$$\begin{aligned}
T_{i2} &= HI_{i2}'\gamma_1 + HI_{i1}^O\gamma_2^O + SI_{i2}'\phi_1 + SY_{i2}\alpha_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f_i'\lambda_{(2)} + \\
&\quad + HI_{i1}^U\gamma_2^U + \mu_{i0}\beta_{(2)} + \varepsilon_{i2} \\
&= HI_{i2}'\gamma_1 + HI_{i1}^O\gamma_2^O + SI_{i2}'\phi_1 + SY_{i2}\alpha_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f_i'\lambda_{(2)} + \\
&\quad + (z_i'\delta + f_i'\kappa + \mathrm{v}_i)\gamma_2^U + \mu_{i0}\beta_{(2)} + \varepsilon_i \\
&= HI_{i2}'\gamma_1 + HI_{i1}^O\gamma_2^O + SI_{i2}'\phi_1 + SY_{i2}\alpha_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f_i'\pi + \\
&\quad + z_i'\psi + e_{H,i}^{CU} \qquad (6)
\end{aligned}$$

where $\pi = \left(\lambda_{(2)} + \kappa\gamma_2^U\right)$, $\psi = \gamma_2^U\delta$ and $e_{H,i}^{CU} = \gamma_2^U v_i + \mu_{i0}\beta_{(2)} + \varepsilon_{i2}$. As discussed in Todd and Wolpin (2007), there should be no correlation between included inputs and the error term v_i for this strategy to overcome the potential omitted variable bias.

3.1.2. The value-added model

The expression given in (5) corresponds to a 'cumulative' model. It is a specification that accounts for all the relevant influences received until the period in which skill is measured. A specification that only accounts for contemporaneous influences and controls for past achievement is typically known as a value-added model. Formally:

$$T_{i2} = \rho T_{i1} + HI_{i2}'\gamma_1 + SI_{i2}'\phi_1 + SY_{i2}\alpha_1 + h_{i2}\varphi_1 + f_i'\lambda + e_{P,i}^{VA} \qquad (7)$$

where $e_{P,i}^{VA} = \mu_{i0}\beta + \varepsilon_{i2} - \rho\varepsilon_{i1}$. This is a restricted version of the cumulative specification, with the restriction being that the effects of *all* Period 1 influences decline at the same rate ρ. To see this, notice that Period 1 skill can be expressed as:

$$T_{i1} = HI_{i1}'\gamma_1 + h_{i1}\varphi_1 + f_i'\lambda + \mu_{i0}\beta + \varepsilon_{i1} \qquad (8)$$

so that:

$$\begin{aligned}
T_{i2} - \rho T_{i1} &= HI_{i2}'\gamma_1 + HI_{i1}'(\gamma_2 - \rho\gamma_1) + SI_{i2}'\phi_1 + SY_{i2}\alpha_1 + h_{i2}\varphi_1 + h_{i1}(\varphi_2 - \rho\varphi_1) + \\
&\quad f_i'\left(\lambda_{(2)} - \rho\lambda\right) + \mu_{i0}\left(\beta_{(2)} - \rho\beta\right) + \varepsilon_{i2} - \rho\varepsilon_{i1} \qquad (9)
\end{aligned}$$

For (9) to reduce to (7) the following must hold: $\gamma_2 = \rho\gamma_1$, $\varphi_2 = \rho\varphi_1$, $\lambda_{(2)} = (1+\rho)\lambda$, and $\beta_{(2)} = (1+\rho)\beta$. This means that the rate of decay of the effects of influences occurring in Period 1 is the same for all influences and equal to ρ[10].

Consistent estimation of the effect of contemporaneous influences requires absence of correlation between the error term $e_{P,i}^{VA}$ and the rest of the right-hand variables in (7). This assumption is particularly problematic for lagged measured skill (T_{i1}) because it will be correlated with its own measurement error unless, as pointed out by Todd and Wolpin (2007), we assume that measurement error is serially correlated with an autoregressive parameter equal to the rate of decay of the effect of inputs. The presence of innate ability in the error term also constitutes a potential source of bias for the OLS estimate of the persistence parameter (ρ) and of input parameters.

[10] Notice that we are allowing predetermined direct influences and innate ability to exert an effect every period equal to λ and β, respectively. If these influences decay at a rate ρ, it means that, by Period 2, the cumulative effect will be given by $\lambda_{(2)} = \lambda + \rho\lambda$ and $\beta_{(2)} = \beta + \rho\beta$, which corresponds to the conditions given above.

An interesting feature of the potential biases affecting the estimation of the persistence parameter is that they operate in opposite directions. As discussed by Andrabi et al. (2011), innate ability will produce a positive bias while measurement error will cause the typical attenuation effect. If we combine this result with the fact that lagged achievement is partially controlling for innate ability, the overall bias affecting ordinary least square (OLS) estimates of contemporaneous input effects might be small.[11]

The inclusion of past inputs in a value-added model allows one to partially relax the assumption that the effects of past inputs share the same rate of decay. This yields:

$$T_{i2} = \rho T_{i1} + HI'_{i2}\gamma_1 + HI^{O}_{i1}\tilde{\gamma}^{O}_2 + SI'_{i2}\phi_1 + SY_{i2}\alpha_1 + h_{i2}\varphi_1 + h_{i1}\tilde{\varphi}_2 + f_i'\lambda + e^{VAP}_{P,i} \quad (10)$$

In an specification like (10) (which will be referred to as the 'value-added-plus' specification following Todd and Wolpin 2007), the effects of included past inputs are allowed to have different rates of decay, which are accommodated through the parameters contained in $\tilde{\gamma}^{O}_2$ and $\tilde{\varphi}_2$. It should be noted that these parameters are not the same as γ^{O}_2 and φ_2 given in (6). In fact, inspection of (9), above, reveals that $\tilde{\gamma}^{O}_2 = \gamma^{O}_2 - \rho\gamma^{O}_1$ and $\tilde{\varphi}_2 = \varphi_2 - \rho\varphi_1$. This provides a simple test for the restriction that the effects of *all* Period 1 influences decline at the same rate ρ, which consists in evaluating the significance of past inputs in a value-added-plus specification. It also shows that in case the restriction fails to hold, parameters $\tilde{\gamma}^{O}_2$ and $\tilde{\varphi}_2$ will accommodate the differences.

3.3. Assigning the contribution of input determinants

The second type of bias mentioned above has received much less attention in the literature. It also has its roots in the presence of omitted direct influences and has to do with the use of predetermined family and child characteristics to control for unobservable inputs. Predetermined family and child characteristics can be proposed for this as long as they refer to skill input determinants – in other words, as long as they are relevant arguments in the demand function of skill inputs. As explained above, doing this configures a 'hybrid' model.

Input determinants can control for several omitted inputs, so one needs to be especially careful when assigning these controls to a particular variable category. A good example of this is family income. Family income is not an input into skill formation, but an input determinant. This means it has an indirect effect on skill through the inputs that are sensitive to families' budget constraints. If family income has a role determining the quantity or quality of inputs received during early childhood and also later at school, and there are unobservable influences related to both environments, it would not be appropriate to assign the contribution of this control exclusively to either of them. Doing so would lead to a biased estimate of the contribution of the category hosting family income.

Based on the above, for this analysis we will assign the contribution of predetermined direct influences and other input determinants to a special category hosting omitted inputs. This is consistent with the notion that family resources and preferences, as well as local educational supply characteristics, play a role in determining not only the educational environment during early childhood $\left(HI'_{i1}\right)$, but also the number of years of schooling (SY^{j}_{i2}) and the educational inputs provided at school (SI^{j}_{i2}) during Period 2. As described in Section 2, a key feature of the Peruvian education system is that school characteristics are highly heterogeneous and

[11] This result is proposed by Andrabi et al. (2011) to explain why OLS estimates of the private school skill premium in Pakistan provided by a valued-added model turned out quite similar to those obtained after applying item response theory and dynamic panel data methods to mitigate the effects of measurement error and observed innate ability.

strongly correlated with children's SES. The use of a special category to host omitted inputs in general will also prevent us from making strong assumptions regarding the nature of unobservable influences.

4. Decomposition strategy

4.1. Empirical specifications and assumptions

The four empirical specifications that stem from combining the hybrid and the value-added options described above are summarised in Table 2.

Table 2. *Empirical specifications*

	Cumulative	Value-added plus
Production function	$T_{i2} = HI_{i2}^{O'}\gamma_1 + HI_{i1}^{O'}\gamma_2 + SI_{i2}^{O'}\phi_1$ $+SY_{i2}\alpha_1 + h_{i2}\varphi_1$ $+h_{i1}\varphi_2 + f_i'\lambda_{(2)}$ $+e_{P,i}^{CU}$	$T_{i2} = \rho T_{i1} + HI_{i2}^{O'}\gamma_1 + HI_{i1}^{O'}\tilde{\gamma}_2$ $+SI_{i2}^{O'}\phi_1 + SY_{i2}\alpha_1$ $+h_{i2}\varphi_1 + h_{i1}\tilde{\varphi}_2 + f_i'\lambda$ $+e_{P,i}^{VAP}$
Hybrid	$T_{i2} = HI_{i2}^{O'}\gamma_1 + HI_{i1}^{O'}\gamma_2 + SI_{i2}^{O'}\phi_1$ $+SY_{i2}\alpha_1 + h_{i2}\varphi_1$ $+h_{i1}\varphi_2 + f_i'\pi + z_i'\psi$ $+e_{H,i}^{CU}$	$T_{i2}\ \rho T_{i1} + HI_{i2}^{O'}\gamma_1 + HI_{i1}^{O'}\tilde{\gamma}_2$ $+SI_{i2}^{O'}\phi_1 + SY_{i2}\alpha_1$ $+h_{i2}\varphi_1 + h_{i1}\tilde{\varphi}_2 + f_i'\tilde{\pi}$ $+ \quad + \quad ,$

The elements contained in the error terms of these specifications play a crucial role in determining the existence of bias in an OLS estimate of the production function parameters. Let us consider a minimum set of assumptions that will serve to impose some structure on these error terms and to identify the individual effects of interest.

(i) The production function of skill can be expressed as a linear function of inputs and predetermined direct influences as follows:

$$A_{i2} = HI_{i2}'\gamma_1 + HI_{i1}'\gamma_2 + SI_{i2}'\phi_1 + SY_{i2}\alpha_1 + h_{i2}\varphi_1 + h_{i1}\varphi_2 + f_i'\lambda_{(2)} + \mu_{i0}\beta_{(2)}$$

$$A_{i1} = HI_{i1}'\gamma_1 + h_{i1}\varphi_1 + f_i'\lambda + \mu_{i0}\beta$$

(ii) Skill is measured with error: $T_{it} = A_{it} + \varepsilon_{it}; t = 1{,}2$.

(iii) Innate ability (μ_{i0}) cannot be observed. Some Period 1 and Period 2 inputs also remain unobserved and are contained in vectors UI_{i1} and UI_{i2}, respectively. Years of schooling, health inputs and predetermined direct influences are observed.

(iv) Skill inputs are determined by parents' choices and it is possible to characterise their demand equations as a linear function of predetermined direct influences (f_i) and other exogenous input determinants (z_i) up to a random error (v_i).

(v) The effects of unobserved Period 1 influences and predetermined direct influences decay at a rate equal to ρ.

Assumptions (i), (ii) and (iii) imply that the error term of the production function cumulative model is given by:

$$e_{P,i}^{CU} = UI_{i1}'\tau_2 + UI_{i2}'\tau_1 + \mu_{i0}\beta_{(2)} + \varepsilon_{i2} \tag{11}$$

where τ_2 and τ_1 contain the production function parameters of Period 1 and Period 2 unobserved inputs.

In addition, assumption (iv) allows one to express the demand functions of unobserved inputs as follows: $UI_{i1} = \delta_1 z_i + \kappa_1 f_i + v_{i1}$ and $UI_{i2} = \delta_2 z_i + \kappa_2 f_i + v_{i2}$[12]. This implies that the error term of the hybrid cumulative model is given by:

$$e_{H,i}^{CU} = v'_{i1}\tau_2 + v'_{i2}\tau_1 + \mu_{i0}\beta_{(2)} + \varepsilon_{i2} \tag{12}$$

and that $\pi = \lambda_{(2)} + \kappa'_1\tau_2 + \kappa'_2\tau_1$ and $\psi = \delta'_1\tau_2 + \delta'_2\tau_1$.

Finally, assumption (v) implies that the error terms of the two value-added-plus specifications can be expressed as:

$$e_{P,i}^{VAP} = UI'_{i2}\tau_1 + \mu_{i0}\big(\beta_{(2)} - \rho\beta\big) + \varepsilon_{i2} - \rho\varepsilon_{i1} \tag{13}$$

$$e_{H,i}^{VAP} = v'_{i2}\tau_1 + \mu_{i0}\big(\beta_{(2)} - \rho\beta\big) + \varepsilon_{i2} - \rho\varepsilon_{i1} \tag{14}$$

and that $\tilde{\pi} = \lambda + \kappa'_2\tau_1$ and $\tilde{\psi} = \delta'_2\tau_1$.

Following the discussion presented in Section 3, the estimation of the production function parameters will be based on the specification that provides the richest set of controls. This is the case of the hybrid value-added-plus model. The reasons for preferring this specification can be summarised as follows: (i) it controls for omitted Period 1 inputs; (ii) it controls for the effect of innate ability accumulated during Period 1; (iii) the remaining (contemporaneous) effect of innate ability will not necessarily bias (upwards) the estimate of the persistence parameter because of the countervailing attenuation effect of measurement error; and (iv) the introduction of exogenous input determinants allows one to control for omitted Period 2 inputs and mitigate omitted variable bias as long as the error terms of their demand functions are not correlated with observed inputs.

Among these features, being able to control for past (Period 1) omitted direct influences is particularly important for obtaining a consistent estimate of the contribution of school inputs. This is because we know that this input category is only relevant in Period 2, so correlation between *any* omitted Period 1 influence and included school inputs would lead to a biased estimate of the contribution of the school influences.

It is worth noting that a value-added model will be able to fully control for innate ability if we assume it only affects initial conditions. We have not imposed this assumption here and, instead, we allow ability to affect the learning process in every period. As a consequence, the error terms of both value-added models include the term $\mu_{i0}\big(\beta_{(2)} - \rho\beta\big)$. This captures the contemporaneous effect of innate ability on skill, and its presence can be of concern for the purpose of identifying the individual effects of included influences and lagged skill. In this setting, the consistency of estimates provided by the value-added models relies on the absence of correlation between the contemporaneous effect of innate ability and included inputs and lagged skill.

In this regard, empirical results presented in several recent studies corroborate the proposition that value-added models can provide reliable estimates of the individual effects of skill inputs. These studies are reviewed in Singh (2015). They show that value-added models such as the one presented in (7) outperform other empirical strategies when recovering teacher effects from simulated data (Guarino et al. 2012), and provide the same results as

[12] Notice that UI_{i1} and UI_{i2} are vectors. Accordingly, each row in the matrices δ_1 and κ_1, and δ_2 and κ_2, contain the parameters of the demand functions of Period 1 and Period 2 unobserved inputs, respectively.

experimental and quasi-experimental methods used to identify school or teacher effects (Deming et al. 2014; Kane et al. 2013, among others). Moreover, value-added estimates given in Singh (2015) for the effect of private school enrolment on the achievement of rural children were also found to be similar to the results provided by an experimental exercise carried out in the same region of India (Muralidharan and Sundararaman 2013).[13]

These results indicate that lagged achievement is a sufficient statistic to control for assignment mechanisms that correlate with ability. This, in terms of the error structures given above, implies that correlation between $\mu_{i0}(\beta_{(2)} - \rho\beta)$ and included inputs is not significant. Correlation between $\mu_{i0}(\beta_{(2)} - \rho\beta)$ and lagged achievement is another potential source of bias, but this is probably being counteracted by the attenuation effect of measurement error, as already pointed out by Andrabi et al. (2011).

The nature of the available data also plays a role in the process of narrowing down a preferred empirical specification. For example, a dataset with a richer set of Period 2 influences than Period 1 influences (like the one described in Section 4.3 below) makes a stronger case for value-added-plus specification.

Preference for the hybrid value-added-plus specification is based on comparing the error structures given in equations (11) to (14), and on the performance of value-added models reported in previous studies. Therefore, the empirical strategy will comprise a comparison of the results obtained in terms of decompositions and parameter estimates across all four specifications. To the extent that these estimates behave in a manner consistent with the error structures given above, we will have more evidence to support our choice of empirical specification.

4.2. Urban–rural gap decompositions

The variables involved in each of the four specifications given in Table 2 will be arranged in different groups or categories. The empirical goal is to measure the contribution of each category to the estimated Period 2 urban–rural gap in cognitive skill, which is given by $\bar{T}_{iU} - \bar{T}_{iR}$ (where an upper bar denotes the sample mean of the variable). This contribution is given by the weighted average of the urban–rural differences in the variables assigned to the category, where the weights are given by estimated values of their production function parameters.

If the specification includes an urban–rural group indicator, the sum of the contributions of all categories will equal the estimated urban–rural gap in cognitive skill. This is because the group indicator ensures that the regression line passes through the means of both groups. Importantly, an urban–rural indicator cannot be regarded as a direct influence but can have a role among input determinants. Accordingly, the inclusion of this indicator will be possible in the hybrid models but not in the production function specifications.

[13] It should be noted that all the empirical exercises cited in this paragraph aimed at obtaining the total (or 'policy') effect of being assigned to a certain type of school input. Nevertheless, assumptions in terms of absence of correlation between the contemporaneous effect of innate ability and included inputs and lagged skill are the same as those required in a value-added model aimed at estimating the direct effects (or production function parameters) of inputs. The difference relies on the controls used in the value-added specification. Total effects require one to control only for the exogenous determinants of excluded inputs, which are the arguments of their conditional demand functions (see Glewwe and Miguel 2008). The estimation of production function parameters requires one to control for the rest of relevant inputs. Hence the variables included in the empirical specifications given in Table 2, and the use of the hybrid specification as a way of controlling for potentially omitted Period 2 influences.

The categories proposed for this analysis have to allow one to identify the contribution of influences originating at school and of influences that occurred during early childhood (Period 1). Categories must also be consistent with the role played by specific variables in the production function of skill. As already discussed, one needs to be especially careful when assigning the contribution of variables that control for omitted inputs. This now becomes clear if we examine the structure behind the parameters associated to exogenous input determinants (z_i) and predetermined direct influences (f_i) in the hybrid models.

Notice that in the hybrid cumulative specification, exogenous input determinants are related to a parameter vector (ψ), which is a function of two sets of parameters: (i) those of exogenous input determinants in the demand functions of omitted inputs (δ_1 and δ_2); and (ii) those of omitted inputs in the production function of skill (τ_2 and τ_1). In fact, $\psi = \delta_1^{'}\tau_2 + \delta_2^{'}\tau_1$.

Predetermined direct influences have both a direct and an indirect effect which operates through the demand for omitted inputs. Their parameters (contained in vector π) are a function of three sets of parameters: (i) the cumulative effect of predetermined direct influences in the production function of skill ($\lambda_{(2)}$); (ii) the parameters of predetermined direct influences in the demand functions of omitted inputs (κ_1 and κ_2); and (iii) the parameters of omitted inputs in the production function of skill (τ_2 and τ_1). In fact, $\pi = \lambda_{(2)} + \kappa_1^{'}\tau_2 + \kappa_2^{'}\tau_1$.

Assumption (v), above, implies that a value-added specification will effectively remove Period 1 unobserved inputs as well as the effect, in Period 1, of predetermined direct influences. Accordingly, in the hybrid value-added-plus model, elements in $\tilde{\psi}$ and $\tilde{\pi}$ will no longer depend on the production or demand function parameters of Period 1 omitted inputs. As indicated above, $\tilde{\pi} = \lambda + \kappa_2^{'}\tau_1$ and $\tilde{\psi} = \delta_2^{'}\tau_1$.

An important implication of the parameter structure described above is that, in absence of further restrictions, it will not be possible to separately identify the direct and indirect effects of predetermined direct influences in a hybrid specification. Strong assumptions are also required to claim that omitted inputs belong only to either the early childhood or school environment.[14] As a consequence, for hybrid specifications we will jointly measure the contribution of predetermined direct influences and all omitted inputs by allocating the contribution of variables contained in f_i and z_i into a special category. 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.

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 fact, rejection of the null hypothesis $(\bar{z}_U - \bar{z}_R)^{'}\psi = 0$ or $(\bar{z}_U - \bar{z}_R)^{'}\tilde{\psi} = 0$ in the corresponding hybrid model implies the presence of at least one omitted input.[15]

[14] The use of school fixed effects would allow one to claim that Period 2 omitted influences belong only to the home environment. School fixed effects, however, can also pick up the effects of Period 1 influences if children are sorted into schools according to ability. As a consequence, school fixed effects typically provide large estimates of the contribution of school variables to achievement gaps (see, for example, McEwan 2004 or McEwan and Trowbridge 2007).

[15] Consider the case of the hybrid-cumulative model. Rejection of the null $(\bar{z}_U - \bar{z}_R)'\psi = 0$ implies that $\psi \neq 0$. Given that $\psi = \delta_1'\tau_2 + \delta_2'\tau_1$, $\psi \neq 0$ implies that either $\delta_1'\tau_2 \neq 0$ or $\delta_2'\tau_1 \neq 0$. It suffices for one of these two inequalities to hold to conclude that there is at least one omitted input and that z_i is a relevant argument in its demand function (i.e. either $\tau_2 \neq 0$ and $\delta_1 \neq 0$ or $\tau_1 \neq 0$ and $\delta_2 \neq 0$). Notice that failure to reject the hypothesis $(\bar{z}_U - \bar{z}_R)'\psi = 0$ does not directly imply the absence of omitted inputs. For example, parameters in ψ can be zero even if $\tau_2 \neq 0$ and $\tau_1 \neq 0$ (i.e. there are omitted inputs from both periods) if the proposed exogenous input determinants have no role in their demand equations ($\delta_1 = 0$ and $\delta_2 = 0$). This could be the case if the variables considered as predetermined direct influences (f_i) fully characterise the demand for omitted inputs.

Rejection of the null $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$ in the hybrid value-added-plus specification further implies the presence of at least one Period 2 omitted input.

It is also worth noting that in the value-added specifications considered we are not restricting the effect of observed inputs to decay at the same rate ρ (hence the name 'value-added plus'). As discussed in the previous section, despite including Period 1 observed inputs, these value-added models will not allow one to recover their direct effects. In fact, the contributions of lagged achievement and Period 1 inputs will have to be merged into a single category of 'past influences'.

The categories related to the four specifications given in Table 2 are summarised in Table 3.

Table 3. *Variable categories related to each empirical specification and their contribution to the urban–rural gap in cognitive skill*

	Cumulative		Value-added plus	
Prod. function	Early childhood and home inputs	$(\overline{HI}^O_{U2} - \overline{HI}^O_{R2})'\hat{\gamma}_1$ $+(\overline{HI}^O_{U1} - \overline{HI}^O_{U1})'\hat{\gamma}_2$	Period 2 home inputs	$(\overline{HI}^O_{U2} - \overline{HI}^O_{R2})'\hat{\gamma}_1$
	Health inputs	$(\bar{h}_{U2} - \bar{h}_{R2})\hat{\varphi}_1$ $+(\bar{h}_{U1} - \bar{h}_{R1})\hat{\varphi}_2$	Period 2 health inputs	$(\bar{h}_{U2} - \bar{h}_{R2})\hat{\varphi}_1$
	School inputs	$(\overline{SI}^O_{U2} - \overline{SI}^O_{R2})'\hat{\phi}_1$ $+(\overline{SY}_{U2} - \overline{SY}_{R2})'\hat{\alpha}_1$	School inputs	$(\overline{SI}^O_{U2} - \overline{SI}^O_{R2})'\hat{\phi}_1$ $+(\overline{SY}_{U2} - \overline{SY}_{R2})'\hat{\alpha}_1$
	Predetermined direct influences	$(\bar{f}_U - \bar{f}_R)'\hat{\lambda}_{(2)}$	Period 2 predetermined direct influences	$(\bar{f}_U - \bar{f}_R)'\hat{\lambda}$
	--	--	Past influences	$\hat{\rho}(\bar{T}_{U1} - \bar{T}_{U2})$ $+(\overline{HI}^O_{U1} - \overline{HI}^O_{U1})'\hat{\tilde{\gamma}}_2$ $+(\bar{h}_{U1} - \bar{h}_{R1})\hat{\tilde{\varphi}}_2$
Hybrid	Early childhood and home inputs	$(\overline{HI}^O_{U2} - \overline{HI}^O_{R2})'\hat{\gamma}_1$ $+(\overline{HI}^O_{U1} - \overline{HI}^O_{U1})'\hat{\gamma}_2$	Period 2 home inputs	$(\overline{HI}^O_{U2} - \overline{HI}^O_{R2})'\hat{\gamma}_1$
	Health inputs	$(\bar{h}_{U2} - \bar{h}_{R2})\hat{\varphi}_1$ $+(\bar{h}_{U1} - \bar{h}_{R1})\hat{\varphi}_2$	Period 2 health inputs	$(\bar{h}_{U2} - \bar{h}_{R2})\hat{\varphi}_1$
	School inputs	$(\overline{SI}^O_{U2} - \overline{SI}^O_{R2})'\hat{\phi}_1$ $+(\overline{SY}_{U2} - \overline{SY}_{R2})'\hat{\alpha}_1$	School inputs	$(\overline{SI}^O_{U2} - \overline{SI}^O_{R2})'\hat{\phi}_1$ $+(\overline{SY}_{U2} - \overline{SY}_{R2})'\hat{\alpha}_1$
	Predetermined direct influences and omitted inputs	$(\bar{f}_U - \bar{f}_R)'\hat{\pi} + (\bar{z}_U - \bar{z}_R)'\hat{\psi}$	Period 2 predetermined direct influences and Period 2 omitted inputs	$(\bar{f}_U - \bar{f}_R)'\hat{\tilde{\pi}} + (\bar{z}_U - \bar{z}_R)'\hat{\tilde{\psi}}$
	--	--	Past influences	$\hat{\rho}(\bar{T}_{U1} - \bar{T}_{U2})$ $+(\overline{HI}^O_{U1} - \overline{HI}^O_{U1})'\hat{\tilde{\gamma}}_2$ $+(\bar{h}_{U1} - \bar{h}_{R1})\hat{\tilde{\varphi}}_2$

Finally, it is worth noting that in the hybrid value-added-plus specification, the contribution of direct school influences will be given by the category labelled 'school inputs' in Table 3. The contribution of direct influences that occurred during early childhood will be estimated through the category labelled 'past influences'. This category includes the contribution of educational and health inputs provided in Period 1, and also the contribution of predetermined direct influences and innate ability that is due to their Period 1 effects (which are subsumed in lagged achievement).

4.3. Data sources and variables

The data for this analysis will be provided by the Peruvian dataset of the Young Lives study. Young Lives is an international study of childhood poverty following 12,000 children in four countries (Ethiopia, India, Peru and Vietnam) over 15 years. In particular, it will consider the first three rounds of the child and household surveys, as well as the school survey, focusing on the Younger Cohort. Table 4 summarises the basic time structure of these data.

Table 4. *Time structure and sample sizes of the relevant Young Lives databases*

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

Source: Young Lives (Peru).

All the information was merged into a single dataset at the child level. Minimum information requirements for inclusion were: (i) having cognitive skill scores for Rounds 2 and 3; and (ii) attending a school included in the school survey.[16] This produced a sample of 487 children in 124 schools.

Consistent with the production function described in Section 3, Period 1 variables will correspond to influences relevant from birth up to the age of 5, and Period 2 variables will correspond to influences relevant between the ages of 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 (i.e. we are assuming that school characteristics have not changed significantly during the two-year period that separates Round 3 from the school survey). We are also assuming that the child has remained in the same school since her enrolment in Grade 1 (at the age of 6) until the school survey was conducted (when she was 10 years old).

Rounds 2 and 3 of the Peruvian Young Lives Younger Cohort child survey contain several measures of cognitive achievement: the Cognitive Development Assessment (CDA), the PPVT, the Early Grade Reading Assessment (EGRA), a maths test, and single items in reading, writing and multiplication. This analysis will focus on the results obtained in the 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.

[16] The risk of selection bias due to this second condition is very small. Primary school attendance in Peru is close to 100 per cent (only 0.7 per cent of Young Lives Younger Cohort children were not attending school in Round 3). In addition, schools participating in the school survey were randomly selected within the four strata considered by the designers of the survey (urban-private, urban-public, rural-public, rural-bilingual-public; see Guerrero et al. 2012). Even so, the second requirement for inclusion will be relaxed in some of the specifications that follow in order to explore if results are affected by the fact of working with children whose school participated in the school survey.

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 5 presents the variables from the child and household surveys considered as early childhood and home inputs, health inputs, predetermined direct influences, and input determinants.

Regarding early childhood and home inputs, the information on household expenditure allows one to approximate the flows invested in the child in Rounds 2 and 3. Since these inputs refer to those with an educational nature, the expenditure flows considered are those that account for learning materials and entertainment. For Period 1, it is possible to approximate expenditure flows invested in the child and to account for access to pre-school by the age of 5 (Round 2). The survey, however, is not very informative of the quality of care and the home environment during infancy (Round 1). Mothers' access to antenatal care (where advice on parenting practices is usually provided) and the way mothers responded to the child crying were used to approximate early educational inputs or the quality of early stimulation provided to the child at home.

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

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

Unlike the educational home inputs, the number of school characteristics that could be potentially considered within the school inputs category is notably large. In Guerrero et al. (2012), the researchers who designed and implemented the school survey distinguish between school-quality and school-responsiveness variables when describing the 'educational opportunities' offered to children at school. Within the school-quality group, variables that are potentially relevant for this analysis can be further classified into five categories: (i) size, organisation and timetable (ORG); (ii) infrastructure (INF); (iii) climate (CLIM); (iv) activities and materials (ACT); and (v) teacher characteristics (TEA).

The first category includes variables capturing the number of students, teachers and classes, the presence of support staff, the number of shifts and teaching hours per day, and teacher attendance. Infrastructure comprises access to basic services and learning facilities. Variables in the school and class climate 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. The activities and materials category comprises information regarding the use of specific learning materials in class and the degree of curriculum coverage attained. Finally, teacher characteristics include teachers' qualifications, experience, and a measure of their 'pedagogical content knowledge' (PCK).[17]

[17] Pedagogical content knowledge as measured in the school survey comprises knowledge of student conceptions and misconceptions about the subject matter (Guerrero et al. 2012). This was measured through teacher responses to questions regarding the sources of error in hypothetical student answers to maths questions.

School responsiveness is related to the degree with which schools respond to students' needs and potential. Variables within this category (RESP) indicate whether or not the school provides support for students lagging behind or at risk of dropping out.

The school survey contains several candidate variables within each of the six categories described above. Because of this, the following three-step procedure was followed to select the variables included as school inputs: (i) pairwise correlations between candidate variables within each category were evaluated, variables with correlation coefficients below 0.6 were chosen and those with a correlation above 0.6 with two or more others were discarded; (ii) a regression of PPVT scores on the variables chosen after (i) was run for each category, and variables with a significant partial correlation were chosen; and (iii) a regression of PPVT scores on the variables chosen after (ii) was run, and those with a significant partial correlation were chosen.[18] Variables presented in Table 5 within the school inputs group are the ones which resulted from this procedure.

Predetermined direct influences refer to variables that are outside the current choice set of families and that have a direct effect on skill. These include caregiver´s educational attainment and age,[19] and child characteristics such as gender, age and mother tongue. These child characteristics can accommodate biological differences between children as well as potential advantages that could be present when solving the test used to measure cognitive skill.[20]

Finally, exogenous input determinants include variables reflecting family resources, parental preferences regarding the child's educational attainment, child and sibling characteristics that can affect parental investments, and an urban/rural indicator. This last variable is intended to capture differences in the general health environment and the availability of educational goods and services in the two geographical domains.

[18] The objective was to reduce the amount of noise within the school inputs category while sacrificing neither the dimensions of school quality considered in the school survey nor the possibility of identifying particular inputs as relevant influences. The results reported in the next section are robust to using the first two principal components obtained from each of the six categories (which explain between 52 per cent and 80 per cent of the variance).

[19] Caregiver engagement in educational activities will likely have a different outcome in terms of child skill formation, depending on the caregiver's parenting skills.

[20] This means that these variables appear as relevant influences when moving from the theoretical relation which explains A_{i2} to the empirical specification for T_{i2}. For example, the PPVT being a vocabulary test, children who do not have Spanish as their mother tongue can face a disadvantage. Differences in cognitive skill driven by some child characteristics (e.g. gender or age) can also be due to differences in unobserved inputs operating through parental preferences. This is the indirect effect of predetermined direct influences discussed above and will be captured within the contribution of the 'predetermined direct influences and omitted inputs' category.

Table 5. *Description of the variables used in the empirical specifications*

Variable type	Variable used in empirical specifications	Database
Period 1 measured cognitive skill (T_{i1})	Standardised raw PPVT score	Round 2
Period 2 measured cognitive skill (T_{i2})	Standardised raw PPVT score[a]	Round 3
Early childhood educational inputs (HI^{o}_{i1})	Real expenditure on 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 crying was affectionate (yes = 1)[b]	Round 1
	Child attended formal pre-school (yes = 1)	Round 2
Period 2 educational home inputs (HI^{o}_{i2})	Real expenditure on child (learning materials and entertainment; x1,000 soles; 2006 prices in urban Lima)	Round 3
	Household had books and child was encouraged to read (yes = 1)	Round 3
	Household had a computer (yes = 1)	Round 3
	Child received help from parents when doing homework (yes = 1)	Round 3
	Hours in a typical day the child spent playing	Round 3
	Hours in a typical day the child spent sleeping	Round 3
	Hours in a typical day the child spent studying	Round 3
Period 1 health input (h_{i1})	Child was stunted (yes = 1)[(c)]	Round 2
Period 2 health input (h_{i2})	Child was stunted (yes = 1)	Round 3
School inputs (SI^{o}_{i2}, SY^{j}_{i2})	Years of schooling (basic education)	Round 3
	Hours in a typical day the child spent at school[d]	Round 3
	CLIM: absence of problems in class (score 12-48)[e]	School survey
	INF: school had basic services (yes = 1)[f]	School survey
	ACT: average curricular coverage in maths and language (average % of topics covered in depth)[e]	School survey
	ORG: teacher absenteeism (%)[g]	School survey
	ORG: school had a psychologist (yes = 1)	School survey
	ORG: school was 'multigrade' (yes = 1)[h]	School survey
	TEA: more than 50% of teachers graduated from a university (yes = 1)[e]	School survey
	TEA: maths teacher's pedagogical content knowledge (score 0–14)	School survey
Predetermined direct influences (f_i)	Child's caregiver had higher education (yes = 1)	Round 3
	Caregiver's age	Round 3
	Child is male (yes = 1)	Round 3
	Child's mother tongue is Spanish (yes = 1)	Round 3
	Child's age in months	Round 3
Exogenous input determinants (z_i)	Child lived 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 was university education (yes=1)	Rounds 2 and 3

a Round 3 and Round 2 raw PPVT scores were standardised using the Round 2 mean and standard deviation.

b Mother cuddled or soothed child when he/she cried.

c A child was considered stunted if she exhibited 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 from the same teacher.

Table 6 presents descriptive statistics as well as urban/rural differences for all the variables described above. Significant differences between urban and rural children are present in most of the direct influences and input determinants considered. Children in the urban domain enjoyed larger expenditure flows from their households, had more access to pre-school and had more opportunities for learning activities at home. In addition, they attended schools with better infrastructure, more qualified teachers and greater curriculum coverage.[21] Finally, urban children also had better health status (were less likely to be stunted) and had more educated caregivers.

These results corroborate that the uneven distribution of direct influences affecting the skill formation process is pervasive across inputs related to different environments and originated at different periods. This, in turn, confirms that the analysis of the relative importance of different input categories when explaining unequal skill outcomes between urban and rural children is far from trivial.

Table 6. *Descriptive statistics and urban–rural gaps*

	Mean	SD	Urban	Rural	Diff.
Standardised raw PPVT score (Round 3)	1.780	0.951	2.095	1.028	1.067*** (0.14)
Standardised raw PPVT score (Round 2)	0.024	0.968	0.355	-0.766	1.121*** (0.13)
Real expenditure on child (learning materials and entertainment; Round 2)[a]	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 crying was affectionate (yes = 1)	0.230	0.421	0.286	0.097	0.188*** (0.05)
Child attended formal pre-school (yes = 1)	0.766	0.424	0.892	0.465	0.427*** (0.055)
Household had books and child was encouraged to read (yes = 1)	0.450	0.498	0.478	0.382	0.096 (0.06)
Household had a computer (yes = 1)	0.140	0.347	0.195	0.007	0.188*** (0.039)
Real expenditure on child (learning materials and entertainment; Round 3)[a]	0.432	0.572	0.517	0.230	0.287*** (0.063)
Child received 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 spent playing	4.346	1.517	4.488	4.005	0.483** (0.218)
Hours in a typical day the child spent sleeping	9.931	0.978	9.988	9.796	0.192 (0.114)
Hours in a typical day the child spent studying	1.945	0.834	2.120	1.526	0.594*** (0.078)
Child was stunted (yes = 1; Round 2)	0.316	0.465	0.207	0.576	-0.369*** (0.034)
Child was stunted (yes = 1; Round 3)	0.189	0.392	0.120	0.354	-0.235*** (0.041)

Continued overleaf

[21] Urban and rural averages reported in Table 6 should not be interpreted as characterising the average urban and rural school. The database is at the child level so figures reported in the above-mentioned table correspond to the ‘average school’ attended by Young Lives children living in urban and rural areas. Characteristics of those schools with more Young Lives children enrolled are given more weight when characterising this ‘average school’.

Table 6. *Descriptive statistics and urban–rural gaps*

	Mean	SD	Urban	Rural	Diff.
Hours in a typical day the child spent 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)
INF: school had 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 had a psychologist (yes = 1)	0.179	0.383	0.248	0.014	0.234* (0.109)
ORG: school was '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)
TEA: teacher's pedagogical content knowledge (PCK score 0–14)	7.587	2.135	8.186	6.168	2.018*** (0.336)
Child's caregiver had 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 lived in urban area (yes = 1)	0.704	0.457	1.000	0.000	1.000
Average household total income[a]	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 was 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 except PCK which has 450.

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

Robust standard errors in parentheses.

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

5. Results and discussion

In this section we present and discuss the results obtained after estimating the contributions presented in Table 3, using the specifications and data described in Tables 2 and 5, respectively.

Table 7 presents normalised contributions (i.e. contributions divided by the urban–rural gap in cognitive skill at the age of 8) and their standard errors. Figure 1 shows the same point estimates accompanied by 95 per cent confidence intervals. Panels (B) and (D) in Figure 1, also present 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. Table A1 in the appendix presents coefficient estimates for the variables involved in all four specifications.

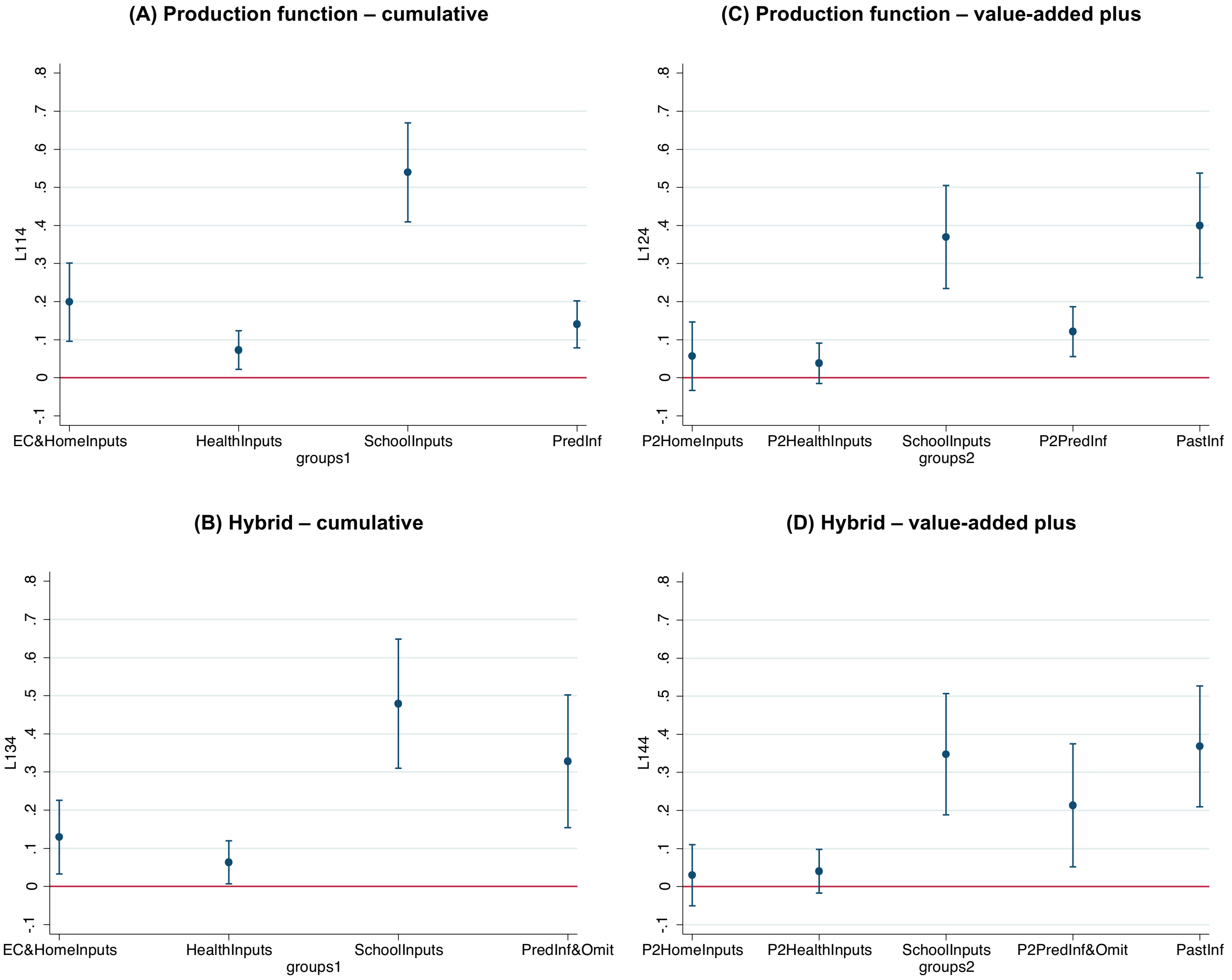
Table 7. *Normalised contributions to the urban–rural gap in cognitive skill at age 8 (% of urban–rural gap)*

	Cumulative (CU)		Value-added plus (VAP)	
Prod. function	Early childhood and home inputs	0.199*** (0.053)	Period 2 home inputs	0.057 (0.046)
	Health inputs	0.073*** (0.026)	Period 2 health inputs	0.038 (0.027)
	School inputs	0.540*** (0.066)	School inputs	0.370*** (0.069)
	Predetermined direct influences	0.141*** (0.032)	Period 2 predetermined direct influences	0.122*** (0.033)
	–	–	Past influences	0.400*** (0.07)
Hybrid	Early childhood and home inputs	0.129*** (0.049)	Period 2 home inputs	0.030 (0.041)
	Health inputs	0.063** (0.029)	Period 2 health inputs	0.041 (0.029)
	School inputs	0.479*** (0.086)	School inputs	0.348*** (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)
	–	–	Past influences	0.368*** (0.081)

Robust standard errors in parentheses.

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

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



(A) Production function – cumulative

(C) Production function – value-added plus

(B) Hybrid – cumulative

(D) Hybrid – value-added plus

Ho: $(\bar{z}_U - \bar{z}_R)'\psi = 0$; stat = 0.22, p-value = 0.031

Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.11, p-value = 0.266

As already discussed, the hybrid value-added-plus model stands out as the preferred specification based on the assumptions and error structures presented in Section 4.1. We will, therefore, consider the results provided by all four specifications. This can be useful in two ways. First, it will allow one to determine if the decompositions behave in a way that is consistent with the reasons for preferring, a priori, the hybrid value-added-plus specification. These reasons refer to the presence of omitted inputs that can bias decomposition results and to the possibility of partially controlling for these unobservable influences using observable exogenous input determinants and lagged skill. Second, it can provide additional insights about the nature of omitted inputs.

Let us start with the results provided by the hybrid cumulative specification (see Panel B in Figure 1). School inputs stand out, with a contribution of nearly 48 per cent of the gap, followed by the category capturing the contribution of predetermined direct influences and omitted inputs (33 per cent). In fact, rejection of the hypothesis $(\bar{z}_U - \bar{z}_R)'\psi = 0$ indicates the presence of omitted inputs and means that exogenous input determinants included in this model have a significant contribution (22 per cent of the gap; $p < 0.05$).

Decomposition results provided by the production function – cumulative specification (Panel A in Figure 1) show somewhat larger contributions for the categories grouping early childhood and home inputs, and school inputs. This suggests that included school inputs can partially pick up the contribution of the omitted influences so, if these do not belong to the school environment, the estimate provided by the production function cumulative specification will be overstating the contribution of school influences. The hybrid cumulative specification has the potential of mitigating this bias[22] but it retains the full (cumulative) effect of unobserved innate ability, another potential source of positive bias in the estimated contribution of inputs determined by families' choices.

Let us now turn to the results provided by the value-added-plus models. As already discussed, these specifications can control for Period 1 omitted inputs and partially control for innate ability.[23] The first thing worth noticing is that in this hybrid specification it is no longer possible to reject the hypothesis that exogenous input determinants have no significant contribution (see Panel D in Figure 1).

Failure to reject the null $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$ does not directly imply the absence of omitted inputs. However, we can combine this result with the fact that: (i) the lack of significance of exogenous input determinants in the hybrid value-added-plus model is not because predetermined direct influences are a sufficient statistic to characterise the demand equation of omitted inputs[24] as these input determinants have a significant contribution in the hybrid cumulative model; and (ii) the inclusion of lagged cognitive test scores causes an important reduction in the contribution of school inputs (compare panels A and C, and B and D in Figure 1) by reducing the estimated effect of several school influences (compare columns 1 and 2, and 3 and 4 in Table A1 in the appendix).

Put together, all this evidence suggests: (i) that there are Period 1 omitted inputs biasing the estimate of the contribution of school influences provided by the cumulative models; and (ii) that there are no significant Period 2 omitted inputs (i.e. $\tau_1 = 0$ in equations (11) to (14)).

It is also worth noting that these results are consistent with the characteristics of the available data, which comprise rich information on Period 2 home and school influences but lack direct information on the quality of the early home environment.

In the appendix (Table A2) we also present decomposition results and parameter estimates obtained when teacher PCK is accounted for among school inputs. This analysis has been done separately because not every school had teachers who participated in the PCK evaluation so the inclusion of this variable implied losing 37 observations (8 per cent of the sample). Inspection of Figure A1 in the appendix reveals that all previous results are robust to the inclusion of teacher PCK and the subsequent reduction of sample size. Apart from the loss in precision, the contribution of school influences remains between 35 per cent and 40 per cent according to the value-added-plus models.

Summing up the evidence, results from the different decompositions indicate the presence of omitted inputs that can be controlled for using available exogenous input determinants and

[22] Something that depends on the degree of correlation between non-observable components of input demand functions.

[23] It is worth noting that significant coefficient estimates for Period 1 influences in the valued-added plus specifications (see columns (2) and (4) in Table A1 in the appendix) reject the restriction that the effect of all Period 1 influences decay at the same rate as the persistence parameter.

[24] If predetermined direct influences are a sufficient statistic to characterise the demand function of omitted inputs, it is possible to accept the null $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$ despite the presence of omitted inputs.

lagged skill. In addition, results suggest that these omitted inputs are primordially related to the early childhood environment. All this evidence is consistent with the assumptions that allow the hybrid value-added-plus model to produce unbiased estimates of individual effects.

Before concluding, it is worth analysing the results obtained after excluding the school inputs contained in the school survey. This implies that the only school inputs considered are years of schooling and time spent at school. This exercise offers an additional opportunity to review the consistency of our results with respect to the assumptions that allow us to identify the contributions of interest, as well as the possibility to check their robustness to consideration of the complete sample of children who have PPVT scores in Rounds 2 and 3. Results are presented in Figures 2 and 3 and coefficient estimates in Tables A3 and A4 in the appendix.

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 in the two hybrid specifications. Consistent with this, the contribution of this category in the hybrid cumulative specification grows twice as large when school survey information is omitted (from 33 per cent in the original decomposition up to 65 per cent according to Panel B in Figure 2). 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 per cent (it was 22 per cent in the original decomposition) and highly significant ($p < 0.00$).

A value-added-plus specification allows a reduction in the contribution of the category hosting omitted inputs (down to 42 per cent; see Panel (D) in Figure 2) and this is consistent with the presence of omitted Period 1 influences, something that has already been suggested by the preceding analysis. Importantly, however, the contribution of exogenous input determinants in the hybrid value-added-plus decomposition remains significant (30 per cent; $p < 0.00$) which implies we cannot accept the null $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$. 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.

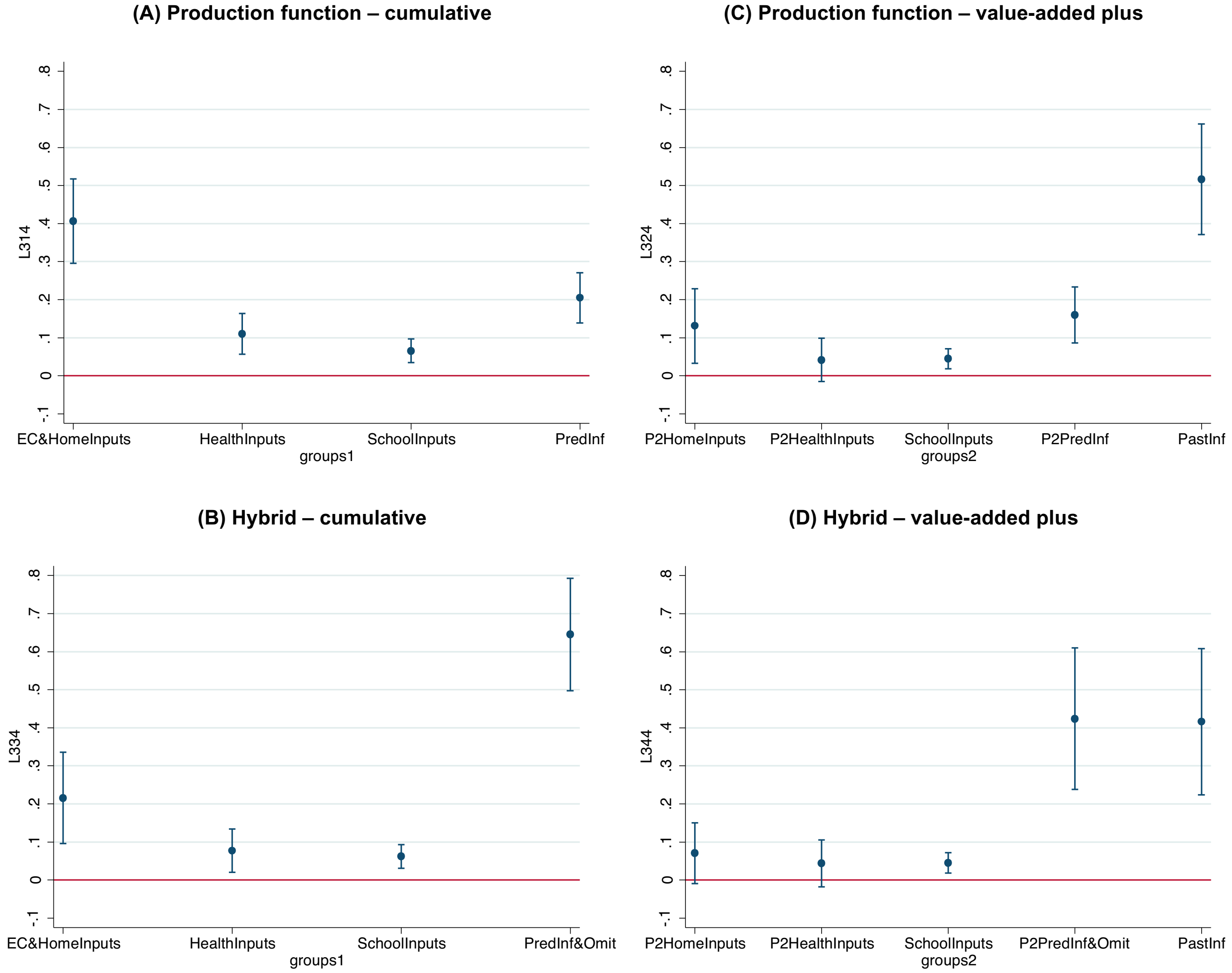
It is also worth noting that the production function cumulative specification provides now a much larger estimate of the contribution of early childhood and home inputs than when considering the full set of variables (40 per cent vs. 20 per cent), and that the school inputs considered in these set of results only account for around 7 per cent of the gap (see Panel A in Figure 2). This means that the school inputs considered are failing to pick up the contribution of their omitted counterparts. This is consistent with the fact that years of schooling and time spent at school are fairly similar across children (between two and three years and around six hours, respectively) and, thus, are weakly correlated with the characteristics of schools and teachers we are omitting.[25]

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

[25] This is just another expression of the fact that, in Peru, relevant inequalities in education are in terms of quality and not in terms of access.

Based on all the above, we can confidently conclude that between 35 and 40 per cent of the gap in cognitive skill between urban and rural 8-year-old children in Peru is related to differences in school inputs (years of schooling, school and teacher characteristics) received between the ages of 6 and 8. This contribution is comparable to that of 'past influences' mostly related to the learning and care environment to which the child was exposed up until the age of 5. At least half of the remaining 20 per cent of the gap can be explained by predetermined direct influences which comprise exogenous child and caregiver characteristics.

Figure 2. *Normalised contributions to the urban–rural gap in cognitive skill at age 8 (point estimates and 95% confidence intervals) – excluding school inputs from the school survey*



(A) Production function – cumulative

(C) Production function – value-added plus

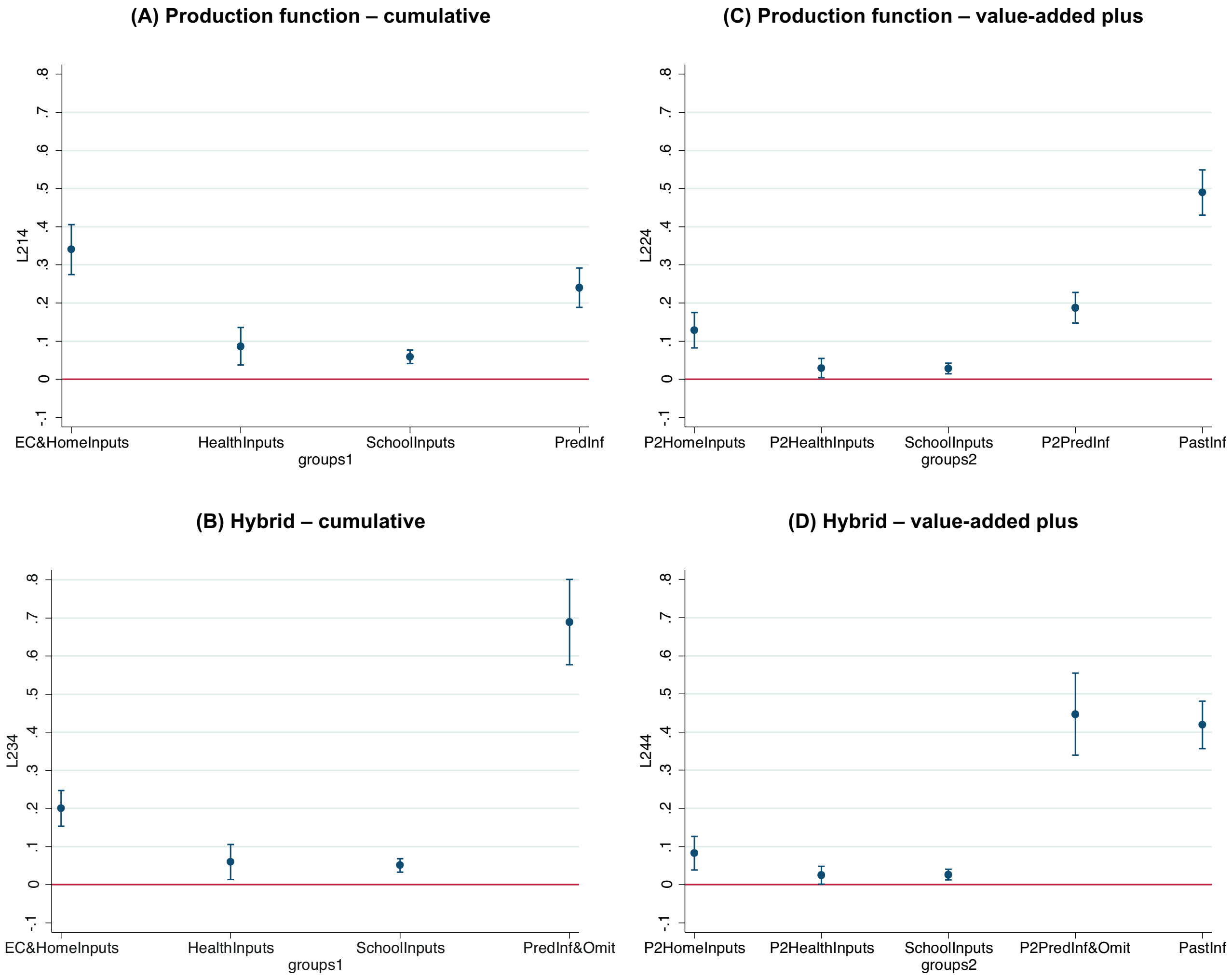
(B) Hybrid – cumulative

(D) Hybrid – value-added plus

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

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

Figure 3. *Normalised contributions to the urban–rural gap in cognitive skill at age 8 (point estimates and 95% confidence intervals) – excluding school inputs from the school survey and working with the complete sample*



(A) Production function – cumulative

(C) Production function – value-added plus

(B) Hybrid – cumulative

(D) Hybrid – value-added plus

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

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

6. Concluding remarks

This analysis aimed at measuring the contribution of school and early childhood influences to the difference in cognitive development observed, at the age of 8, between urban and rural children in Peru.

Empirical work that has analysed the cognitive skill gap between indigenous and non-indigenous Peruvian children has found that family variables contribute more than school variables (Hernandez-Zavala et al. 2006) and that differences in child and household characteristics contribute significantly more than differences in community-level variables (Arteaga and Glewwe 2014). In addition, evidence favouring the role of household characteristics also seems consistent with the fact that setbacks in cognitive development affecting children from disadvantaged backgrounds emerge during early childhood and remain fairly unchanged once these children enter school.

This analysis has contributed new evidence based on an unusually rich dataset and provided by a decomposition strategy less prone to biases than those used so far in the literature. Two key features of the framework that allow it to be less susceptible to biased decomposition results are the explicit distinction between skill inputs and skill input determinants, and the use of a special category to host omitted influences. As discussed in the preceding analysis, including skill input determinants to control for omitted influences can be useful when decomposing a gap, but only as long as we are aware of the assumptions we are making when we assign these controls to a particular category. If family resources and preferences, and local supply characteristics determine the quantity and quality of inputs received both at home and at school, these controls will capture the contribution of inputs that belong to both environments, so a bias can arise if we assign them exclusively to one of them.

The focus on skill inputs is another important feature of this analysis which allows one to draw useful policy implications. This is because inputs, by definition, have a direct effect on skill. If the supply of inputs found to have a significant contribution to some form of inequality can be directly affected by policy action, then policy can have an immediate role in mitigating this inequality.

Contrast this to an analysis focused on input determinants. Because input determinants at the household level (such as family resources or preferences) can affect a wide variety of inputs, they will appear as significant contributors to a gap once inputs are omitted (either purposely or because they remain unobserved). The problem with this result is that it can end up obscuring the role of inputs that are directly affected by policy and centring the attention of policymakers on household variables that have a less obvious relationship with policy action or only an indirect effect on skill (such as parental education or family income).

The results obtained in this analysis indicate that school influences occurring between the ages of 6 and 8 account for a share of the urban–rural cognitive skill gap (35–40 per cent) that is similar to the share attributable to differences in these children's early environment (between the ages of 0 and 5). Because this gap in cognitive skill was already present at the age of 5, these results imply that observable differences between rural and urban schools[26] are serving to sustain the inequalities in cognitive development that emerged before children were enrolled in school. As shown by this analysis, the fact that gaps are already present at pre-school age does not necessarily imply that schools play only a minor role. The reason for this is that skill exhibits less-than-perfect persistence so only a fraction of the previous gap remains every period.

The characteristics of rural schools have a direct connection with policy action because nearly all the supply of educational services in rural areas is public. This, together with the results summarised above, lead to a clear policy implication: efforts devoted to the equalisation of the characteristics of rural schools and teachers with those existing in urban areas could potentially allow a significant reduction (of up to 40 per cent) in the cognitive skill gap between urban and rural children by the time they reach Grade 3.

While we do not focus on the individual causal effects of particular school-level inputs in this paper, we identify a number of indicators of school quality which warrant further research in terms of their potential influence in compounding the educational disadvantage of rural pupils, including those relating to school climate and organisation. Examining the qualitative differences in schools which relate to differences on indicators such as the use of multigrade teaching and the incidence of teacher-reported ‘problems in class’, for example, may improve understanding of the differences in learning outcomes between urban and rural schools in the context of Peru.

[26] The urban–rural indicator used in this analysis corresponds to the domain where the child lives. This means that the relevant difference is between schools attended by urban and rural children. Drawing a conclusion in terms of urban and rural schools, however, is acceptable as 95 per cent of rural children in the sample attend a rural school, while 99 per cent of urban children in the sample attend an urban school.

References

Andrabi, T., J. Das, A.I. Khwaja and T. Zajonc (2011) 'Do Value-Added Estimates Add Value? Accounting for Learning Dynamics', *American Economic Journal: Applied Economics* 3: 29–54.

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

Baker, D., B. Goesling and G. LeTendre (2002) 'Socioeconomic Status, School Quality, and National Economic Development: A Cross−National Analysis of the “Heyneman−Loxley Effect” on Mathematics and Science Achievement', *Comparative Education Review* 46: 291–312.

Beltran, A. and J. Seinfeld (2012) *La trampa educativa en el Peru: cuando la educacion llega a muchos pero sirve a pocos*, Lima: Universidad del Pacifico.

Coleman, J., E. Campbell, C. Hobson, J. McPartland, A. Mood, F. Weinfall and R. York (1966) *Equality of Educational Opportunity*, Washington, DC: United States Department of Health, Education, and Welfare.

Crouch, L., M. Gustafsson and P. Lavado (2009) 'Measuring Educational Inequality in South Africa and Peru' in D. Holsinger and W.J. Jacob (eds) *Inequality in Education*, Dordrecht: Springer Netherlands.

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

Cueto, S., G. Guerrero, J. Leon, M. Zapata and S. Freire (2014a) '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., J. Leon and I. Munoz (2014b) 'Education Opportunities and Learning Outcomes of Children in Peru: A Longitudinal Model' in M. Bourdillon and J. Boyden (eds) *Growing Up in Poverty: Findings from Young Lives*: Basingstoke: Palgrave Macmillan.

Cueto, S., G. Guerrero, J. Leon, M. Zapata and S. Freire (2013) *¿La cuna marca las oportunidades y el rendimiento educativo? Una mirada al caso peruano*, Documento de Investigacion 66, Lima: GRADE.

Deming, D., J. Hastings, T. Kane and D. Staiger (2014) 'School Choice, School Quality, and Post-secondary Attainment', *American Economic Review* 104.3: 991–1013.

Dunn, L., E. Padilla, D. Lugo and L. Dunn (1986) *Manual del examinador para el Test de Vocabulario en Imágenes Peabody: adaptación hispanoamericana*, Minnesota, MN: AGS.

Glewwe, P. and E. Miguel (2008) 'The Impact of Child Health and Nutrition on Education in Less Developed Countries' in T.P. Schultz and J.A. Strauss (eds) *Handbook of Development Economics* (Vol. 4), Amsterdam: North-Holland.

Guarino, C., M.D. Reckase and J.M. Wooldridge (2012) *Can Value-Added Measures of Teacher Performance Be Trusted?,* IZA Discussion Paper 6602, Bonn: Institute for the Study of Labor.

Guerrero, G., J. Leon, E. Rosales, M. Zapata, S. Freire, V. Saldarriaga and S. Cueto (2012) *Young Lives School Survey in Peru: Design and Initial Findings*, Working Paper 92, Oxford: Young Lives.

Hanushek, E. and J. Luque (2003) 'Efficiency and Equity in Schools Around the World', *Economics of Education Review* 22: 481–502.

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

Heyneman, S. and W. Loxley (1983) 'The Effect of Primary-School Quality on Academic Achievement Across Twenty-Nine High- and Low-Income Countries', *American Journal of Sociology* 88: 1162–94.

Heyneman, S. and W. Loxley (1982) 'Influences on Academic Achievement Across High and Low Income Countries: A Re-Analysis of IEA Data', *Sociology of Education* 55: 13–21.

INEI (2012) *Informe Tecnico: Evolucion de la Pobreza 2007–2011*, Lima: Instituto Nacional de Estadistica e Informatica, Peru.

Kane, T., D. McCaffrey, T. Miller and D. Staiger (2013) *Have we Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment,* Research Paper, Measures of Effective Teaching Project, http://www.metproject.org/downloads/MET_Validating_Using_Random_Assignment_Research_Paper.pdf (accessed 2 February 2015).

McEwan, P. (2004) 'The Indigenous Test Score Gap in Bolivia and Chile', *Economic Development and Cultural Change* 53: 157–90.

McEwan, P. and M. Trowbridge (2007) 'The Achievement of Indigenous Students in Guatemalan Primary Schools', *International Journal of Educational Development* 27: 61–76.

MINEDU (2013) *¿Cuanto aprenden nuestros niños? Resultados de la Evaluacion Censal de Estudiantes - ECE 2012*, Lima: Ministerio de Educacion, Peru.

Muralidharan, K. and V. Sundararaman (2013) *The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India*, Working Paper 19441, Cambridge, MA: National Bureau of Economic Research.

OECD (2013) *PISA 2012 Results in Focus: What 15-Year-Olds Know and What They Can Do With What They Know*, Paris: OECD Publishing.

Outes-Leon, I., C. Porter and A. Sánchez (2011) *Early Nutrition and Cognition in Peru: A Within-Sibling Investigation*, Working Paper 241, Washington, DC: Inter-American Development Bank.

Peaker, G. (1971) *The Plowden Children Four Years Later*, London: National Foundation for Educational Research in England and Wales.

Rosenzweig, M. and P. Schultz (1983) 'Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs, and their Effects on Birth Weight', *Journal of Political Economy* 91.5: 723–46.

Schady, N., J. Behrman, M.C. Araujo, R. Azuero, R. Bernal, D. Bravo, F. Lopez-Boo, K. Macours, D. Marshall, C. Paxson and R. Vakis (2014) *Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries*, 482, Washington, DC: 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.

Todd, P. and K. Wolpin (2007) 'The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps', *Journal of Human Capital* 1.1: 91–136.

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

World Bank (2014) 'Indicators', World Bank website, http://data.worldbank.org/indicator (accessed 24 November 2014).

World Bank (2007) *Toward High-Quality Education in Peru: Standards, Accountability and Capacity Building*, Washington, DC: World Bank.

Appendix:

Coefficient estimates for the four empirical specifications and decomposition results including teacher pedagogical content knowledge (PCK)

Table A1. *Coefficient estimates for the four specifications (including school inputs from the school survey)*

	(1)	(2)	(3)	(4)
VARIABLES	Prod.Fn-CU	Prod.Fn-VAP	Hybrid-CU	Hybrid-VAP
Real expenditure on child (learning materials and entertainment; Round 2)	0.125	0.021	0.076	-0.022
	(0.095)	(0.086)	(0.090)	(0.083)
Mother had antenatal visits during pregnancy (yes = 1)	0.137**	0.129**	0.132**	0.121**
	(0.055)	(0.052)	(0.051)	(0.052)
Maternal response to child crying was affectionate (yes = 1)	0.086	0.030	0.077	0.025
	(0.064)	(0.070)	(0.063)	(0.070)
Child attended formal pre-school (yes = 1)	0.088	0.020	0.049	0.003
	(0.083)	(0.071)	(0.094)	(0.080)
Household had books and child was encouraged to read (yes = 1)	0.220***	0.242***	0.210***	0.235***
	(0.063)	(0.069)	(0.057)	(0.068)
Household had a computer (yes = 1)	0.106*	0.074	0.087	0.064
	(0.059)	(0.054)	(0.059)	(0.055)
Real expenditure on child (learning materials and entertainment; Round 3)	0.063	0.035	0.062	0.036
	(0.068)	(0.069)	(0.062)	(0.064)
Child received help from parents when doing homework (yes = 1)	0.087	0.022	0.007	-0.041
	(0.098)	(0.097)	(0.093)	(0.091)
Hours in a typical day the child spent playing	0.004	-0.009	-0.003	-0.012
	(0.026)	(0.029)	(0.021)	(0.026)
Hours in a typical day the child spent sleeping	-0.028	-0.044	-0.032	-0.044
	(0.038)	(0.038)	(0.036)	(0.038)
Hours in a typical day the child spent studying	0.062	0.033	0.045	0.024
	(0.046)	(0.040)	(0.045)	(0.041)
Child was stunted (yes = 1; Round 2)	-0.083	-0.026	-0.048	-0.002
	(0.075)	(0.078)	(0.085)	(0.084)
Child was stunted (yes = 1; Round 3)	-0.200*	-0.173	-0.212	-0.184
	(0.112)	(0.123)	(0.123)	(0.133)
Hours in a typical day the child spent at school	-0.075	-0.067	-0.075	-0.070
	(0.049)	(0.051)	(0.051)	(0.053)
Years of schooling (basic education)	0.348***	0.252***	0.342***	0.253***
	(0.082)	(0.077)	(0.081)	(0.075)
CLIM: absence of problems in class (score 12-48)	0.011**	0.012**	0.009**	0.011**
	(0.004)	(0.004)	(0.004)	(0.004)
INF: school had basic services (yes = 1)	0.229***	0.053	0.183**	0.044
	(0.064)	(0.065)	(0.069)	(0.064)

Continued overleaf

Table A1. *Coefficient estimates for the four specifications (including school inputs from the school survey)* continued

	(1)	(2)	(3)	(4)
VARIABLES	**Prod.Fn-CU**	**Prod.Fn-VAP**	**Hybrid-CU**	**Hybrid-VAP**
ACT: average curricular coverage (% of topics covered in depth)	0.612*	0.432	0.521	0.388
	(0.308)	(0.261)	(0.308)	(0.267)
ORG: teacher absenteeism (%)	-0.862*	-0.830**	-0.780*	-0.761**
	(0.416)	(0.323)	(0.380)	(0.317)
ORG: school had a psychologist (yes = 1)	0.173*	0.187*	0.194**	0.203**
	(0.085)	(0.090)	(0.079)	(0.087)
ORG: school was ‘multigrade’ (yes = 1)	-0.320**	-0.326***	-0.292**	-0.308***
	(0.119)	(0.091)	(0.120)	(0.098)
TEA: more than 50% of teachers graduated from university (yes = 1)	0.102**	0.016	0.090*	0.012
	(0.047)	(0.047)	(0.049)	(0.046)
Child’s caregiver had higher education (yes = 1)	0.187**	0.055	0.135*	0.017
	(0.064)	(0.049)	(0.066)	(0.056)
Caregiver’s age	0.000	-0.002	0.013**	0.008
	(0.004)	(0.004)	(0.005)	(0.005)
Child is male (yes = 1)	0.013	-0.018	0.028	0.034
	(0.039)	(0.035)	(0.096)	(0.083)
Child’s mother tongue is Spanish (yes = 1)	0.348***	0.370***	0.341**	0.389**
	(0.113)	(0.120)	(0.144)	(0.152)
Child’s age in months	0.013	0.003	0.014	0.004
	(0.012)	(0.009)	(0.011)	(0.009)
Child lived in urban area (yes = 1)			0.120	0.028
			(0.113)	(0.096)
Average household total income			0.018	0.010
			(0.024)	(0.021)
Average household size			0.013	0.013
			(0.026)	(0.026)
Proportion of male siblings			-0.032	-0.100
			(0.175)	(0.173)
Child birth order			-0.097***	-0.085**
			(0.026)	(0.031)
Caregiver aspiration for child was university education (yes = 1)			0.058	0.026
			(0.054)	(0.059)
Standardised raw PPVT score (Round 2)		0.348***		0.341***
		(0.049)		(0.052)
Constant	-1.259	0.572	-1.362	0.465
	(1.249)	(1.145)	(1.176)	(1.117)
Observations	487	487	487	487
R-squared	0.547	0.601	0.557	0.608

Robust standard errors in parentheses

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

Table A2. *Coefficient estimates for the four specifications (including school inputs from the school survey and teacher PCK)*

	(1)	(2)	(3)	(4)
VARIABLES	Prod.Fn-CU	Prod.Fn-VAP	Hybrid-CU	Hybrid-VAP
Real expenditure on child (learning materials and entertainment; Round 2)	0.162	0.050	0.109	0.007
	(0.103)	(0.090)	(0.102)	(0.086)
Mother had antenatal visits during pregnancy (yes = 1)	0.157**	0.145**	0.147**	0.131**
	(0.063)	(0.060)	(0.058)	(0.060)
Maternal response to child crying was affectionate (yes = 1)	0.104	0.029	0.096	0.027
	(0.064)	(0.065)	(0.069)	(0.067)
Child attended formal pre-school (yes = 1)	0.036	-0.038	-0.008	-0.058
	(0.094)	(0.083)	(0.100)	(0.089)
Household had books and child was encouraged to read (yes = 1)	0.230***	0.244***	0.220***	0.237***
	(0.063)	(0.069)	(0.059)	(0.069)
Household had a computer (yes = 1)	0.104	0.082*	0.084	0.075
	(0.059)	(0.046)	(0.057)	(0.050)
Real expenditure on child (learning materials and entertainment; Round 3)	0.054	0.030	0.055	0.032
	(0.071)	(0.075)	(0.064)	(0.070)
Child received help from parents when doing homework (yes = 1)	0.083	0.023	0.009	-0.034
	(0.104)	(0.100)	(0.102)	(0.097)
Hours in a typical day the child spent playing	-0.004	-0.021	-0.012	-0.023
	(0.029)	(0.031)	(0.025)	(0.027)
Hours in a typical day the child spent sleeping	-0.031	-0.049	-0.033	-0.047
	(0.043)	(0.040)	(0.039)	(0.039)
Hours in a typical day the child spent studying	0.068	0.034	0.051	0.026
	(0.047)	(0.042)	(0.048)	(0.043)
Child was stunted (yes = 1; Round 2)	-0.070	-0.014	-0.037	0.008
	(0.080)	(0.085)	(0.090)	(0.090)
Child was stunted (yes = 1; Round 3)	-0.245*	-0.212	-0.262*	-0.223
	(0.122)	(0.134)	(0.131)	(0.142)
Hours in a typical day the child spent at school	-0.104*	-0.090*	-0.106*	-0.092*
	(0.050)	(0.049)	(0.050)	(0.049)
Years of schooling (basic education)	0.352***	0.265***	0.343***	0.262***
	(0.091)	(0.080)	(0.093)	(0.080)
CLIM: absence of problems in class (score 12-48)	0.015***	0.015***	0.013**	0.014**
	(0.005)	(0.005)	(0.005)	(0.005)
INF: school had basic services (yes = 1)	0.202***	0.023	0.155**	0.012
	(0.054)	(0.058)	(0.066)	(0.060)
ACT: average curricular coverage (% of topics covered in depth)	0.755**	0.499	0.667*	0.461
	(0.347)	(0.282)	(0.337)	(0.282)
ORG: teacher absenteeism (%)	-0.865*	-0.789**	-0.771*	-0.707**
	(0.428)	(0.321)	(0.405)	(0.326)
ORG: school had a psychologist (yes = 1)	0.223***	0.251***	0.238***	0.265***
	(0.067)	(0.068)	(0.068)	(0.073)
ORG: school was 'multigrade' (yes = 1)	-0.357***	-0.370***	-0.327**	-0.348***
	(0.114)	(0.091)	(0.112)	(0.094)

Continued overleaf

Table A2. *Coefficient estimates for the four specifications (including school inputs from the school survey and teacher PCK)* continued

	(1)	(2)	(3)	(4)
VARIABLES	Prod.Fn-CU	Prod.Fn-VAP	Hybrid-CU	Hybrid-VAP
TEA: more than 50% of teachers graduated from university (yes = 1)	0.100*	0.017	0.089	0.013
	(0.046)	(0.047)	(0.052)	(0.048)
TEA: teacher's pedagogical content knowledge (PCK score 0-14)	0.008	0.018	0.007	0.020
	(0.020)	(0.015)	(0.019)	(0.015)
Child's caregiver had higher education (yes = 1)	0.208***	0.042	0.153**	0.001
	(0.058)	(0.059)	(0.062)	(0.063)
Caregiver's age	-0.001	-0.003	0.011**	0.006
	(0.004)	(0.004)	(0.005)	(0.005)
Child is male (yes = 1)	0.032	0.001	0.052	0.068
	(0.043)	(0.040)	(0.102)	(0.088)
Child's mother tongue is Spanish (yes = 1)	0.312**	0.315**	0.305**	0.337**
	(0.110)	(0.119)	(0.139)	(0.147)
Child's age in months	0.014	0.002	0.014	0.002
	(0.011)	(0.010)	(0.011)	(0.009)
Child lived in urban area (yes = 1)			0.119	0.014
			(0.121)	(0.096)
Average household total income			0.012	0.005
			(0.026)	(0.022)
Average household size			0.011	0.010
			(0.027)	(0.028)
Proportion of male siblings			-0.045	-0.140
			(0.188)	(0.181)
Child birth order			-0.088***	-0.074*
			(0.029)	(0.034)
Caregiver aspiration for child was university education (yes = 1)			0.092	0.055
			(0.056)	(0.064)
Standardised raw PPVT score (Round 2)		0.364***		0.360***
		(0.054)		(0.060)
Constant	-1.252	0.745	-1.291	0.636
	(1.158)	(1.122)	(1.107)	(1.091)
Observations	450	450	450	450
R-squared	0.554	0.609	0.563	0.614

Robust standard errors in parentheses

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

Table A3. *Coefficient estimates for the four specifications excluding school inputs from the school survey*

	(1)	(2)	(3)	(4)
VARIABLES	Prod.Fn-CU	Prod.Fn-VAP	Hybrid-CU	Hybrid-VAP
Real expenditure on child (learning materials and entertainment; Round 2)	0.281**	0.099	0.180	0.038
	(0.124)	(0.093)	(0.106)	(0.087)
Mother had antenatal visits during pregnancy (yes = 1)	0.143**	0.129**	0.136**	0.119**
	(0.054)	(0.050)	(0.050)	(0.053)
Maternal response to child crying was affectionate (yes = 1)	0.132*	0.047	0.101	0.035
	(0.071)	(0.071)	(0.062)	(0.065)
Child attended formal pre-school (yes = 1)	0.184	0.078	0.050	0.009
	(0.107)	(0.098)	(0.121)	(0.113)
Household had books and child was encouraged to read (yes = 1)	0.229***	0.258***	0.207***	0.240***
	(0.063)	(0.068)	(0.052)	(0.062)
Household had a computer (yes = 1)	0.195**	0.129*	0.133*	0.099
	(0.079)	(0.072)	(0.075)	(0.072)
Real expenditure on child (learning materials and entertainment; Round 3)	0.088	0.041	0.077	0.041
	(0.069)	(0.070)	(0.058)	(0.064)
Child received help from parents when doing homework (yes = 1)	0.141	0.058	0.027	-0.023
	(0.107)	(0.101)	(0.085)	(0.085)
Hours in a typical day the child spent playing	0.066**	0.030	0.032*	0.015
	(0.023)	(0.028)	(0.017)	(0.022)
Hours in a typical day the child spent sleeping	0.023	-0.012	-0.011	-0.028
	(0.039)	(0.037)	(0.032)	(0.035)
Hours in a typical day the child spent studying	0.154***	0.081*	0.085*	0.047
	(0.043)	(0.042)	(0.046)	(0.046)
Child was stunted (yes = 1; Round 2)	-0.179*	-0.082	-0.079	-0.025
	(0.087)	(0.088)	(0.101)	(0.094)
Child was stunted (yes = 1; Round 3)	-0.219*	-0.190	-0.226	-0.198
	(0.118)	(0.132)	(0.134)	(0.143)
Hours in a typical day the child spent at school	-0.049	-0.045	-0.042	-0.043
	(0.060)	(0.054)	(0.056)	(0.054)
Years of schooling (basic education)	0.342***	0.225***	0.326***	0.227***
	(0.083)	(0.063)	(0.077)	(0.060)
Child's caregiver had higher education (yes = 1)	0.260***	0.075	0.169**	0.025
	(0.081)	(0.060)	(0.075)	(0.058)
Caregiver's age	-0.003	-0.005	0.012**	0.007
	(0.004)	(0.004)	(0.005)	(0.005)
Child is male (yes = 1)	-0.005	-0.037	0.056	0.062
	(0.044)	(0.036)	(0.094)	(0.078)
Child's mother tongue is Spanish (yes = 1)	0.505***	0.479***	0.390**	0.439**
	(0.115)	(0.131)	(0.145)	(0.170)
Child's age in months	0.016	0.005	0.018	0.007
	(0.013)	(0.010)	(0.011)	(0.010)

Continued overleaf

Table A3. *Coefficient estimates for the four specifications excluding school inputs from the school survey* continued

	(1)	(2)	(3)	(4)
VARIABLES	Prod.Fn-CU	Prod.Fn-VAP	Hybrid-CU	Hybrid-VAP
Child lived in urban area (yes = 1)			0.395***	0.202**
			(0.076)	(0.091)
Average household total income			0.033	0.021
			(0.032)	(0.027)
Average household size			0.006	0.006
			(0.028)	(0.028)
Proportion of male siblings			-0.137	-0.202
			(0.146)	(0.138)
Child birth order			-0.101***	-0.088***
			(0.025)	(0.029)
Caregiver aspiration for child was university education (yes = 1)			0.112	0.065
			(0.072)	(0.070)
Standardised raw PPVT score (Round 2)		0.399***		0.363***
		(0.038)		(0.046)
Constant	-2.008	0.228	-1.894	0.102
	(1.347)	(1.255)	(1.152)	(1.172)
Observations	487	487	487	487
R-squared	0.474	0.556	0.510	0.573

Robust standard errors in parentheses

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

Table A4. *Coefficient estimates for the four specifications excluding school inputs from the school survey and working with the complete sample*

	(1)	(2)	(3)	(4)
VARIABLES	Prod.Fn-CU	Prod.Fn-VAP	Hybrid-CU	Hybrid-VAP
Real expenditure on child (learning materials and entertainment; Round 2)	0.170***	0.050	0.122***	0.034
	(0.043)	(0.030)	(0.036)	(0.028)
Mother had antenatal visits during pregnancy (yes = 1)	0.188***	0.134***	0.193***	0.139***
	(0.048)	(0.038)	(0.044)	(0.039)
Maternal response to child crying was affectionate (yes = 1)	0.055	0.015	0.042	0.012
	(0.054)	(0.051)	(0.043)	(0.046)
Child attended formal pre-school (yes = 1)	0.158**	0.062	0.059	0.016
	(0.059)	(0.045)	(0.058)	(0.044)
Household had books and child was encouraged to read (yes = 1)	0.219***	0.216***	0.183***	0.194***
	(0.059)	(0.053)	(0.048)	(0.048)
Household had a computer (yes = 1)	0.226***	0.126***	0.130***	0.082**
	(0.041)	(0.036)	(0.040)	(0.034)
Real expenditure on child (learning materials and entertainment; Round 3)	0.081*	0.037	0.059	0.029
	(0.045)	(0.030)	(0.044)	(0.032)
Child received help from parents when doing homework (yes = 1)	0.165***	0.111**	0.076	0.061
	(0.044)	(0.046)	(0.045)	(0.044)
Hours in a typical day the child spent playing	0.067***	0.035**	0.037*	0.021
	(0.021)	(0.016)	(0.021)	(0.018)
Hours in a typical day the child spent sleeping	0.009	-0.024	-0.019	-0.036
	(0.025)	(0.023)	(0.024)	(0.024)
Hours in a typical day the child spent studying	0.130***	0.044**	0.065***	0.017
	(0.024)	(0.018)	(0.022)	(0.018)
Child was stunted (yes = 1; Round 2)	-0.154**	-0.044	-0.098	-0.024
	(0.070)	(0.050)	(0.066)	(0.050)
Child was stunted (yes = 1; round 3)	-0.164**	-0.131**	-0.125**	-0.111**
	(0.060)	(0.057)	(0.052)	(0.053)
Hours in a typical day the child spent at school	0.036	0.011	0.020	0.005
	(0.035)	(0.033)	(0.037)	(0.034)
Years of schooling (basic education)	0.294***	0.142***	0.254***	0.134***
	(0.045)	(0.035)	(0.044)	(0.034)
Child's caregiver had higher education (yes = 1)	0.308***	0.103*	0.194***	0.054
	(0.061)	(0.052)	(0.058)	(0.048)
Caregiver's age	-0.002	-0.001	0.009**	0.006
	(0.003)	(0.003)	(0.004)	(0.004)
Child is male (yes = 1)	0.061	0.042	0.068	0.060
	(0.036)	(0.029)	(0.059)	(0.038)
Child's mother tongue is Spanish (yes = 1)	0.513***	0.495***	0.278**	0.367***
	(0.076)	(0.064)	(0.105)	(0.100)
Child's age in months	0.010	0.007	0.012	0.008
	(0.008)	(0.006)	(0.007)	(0.006)

Continued overleaf

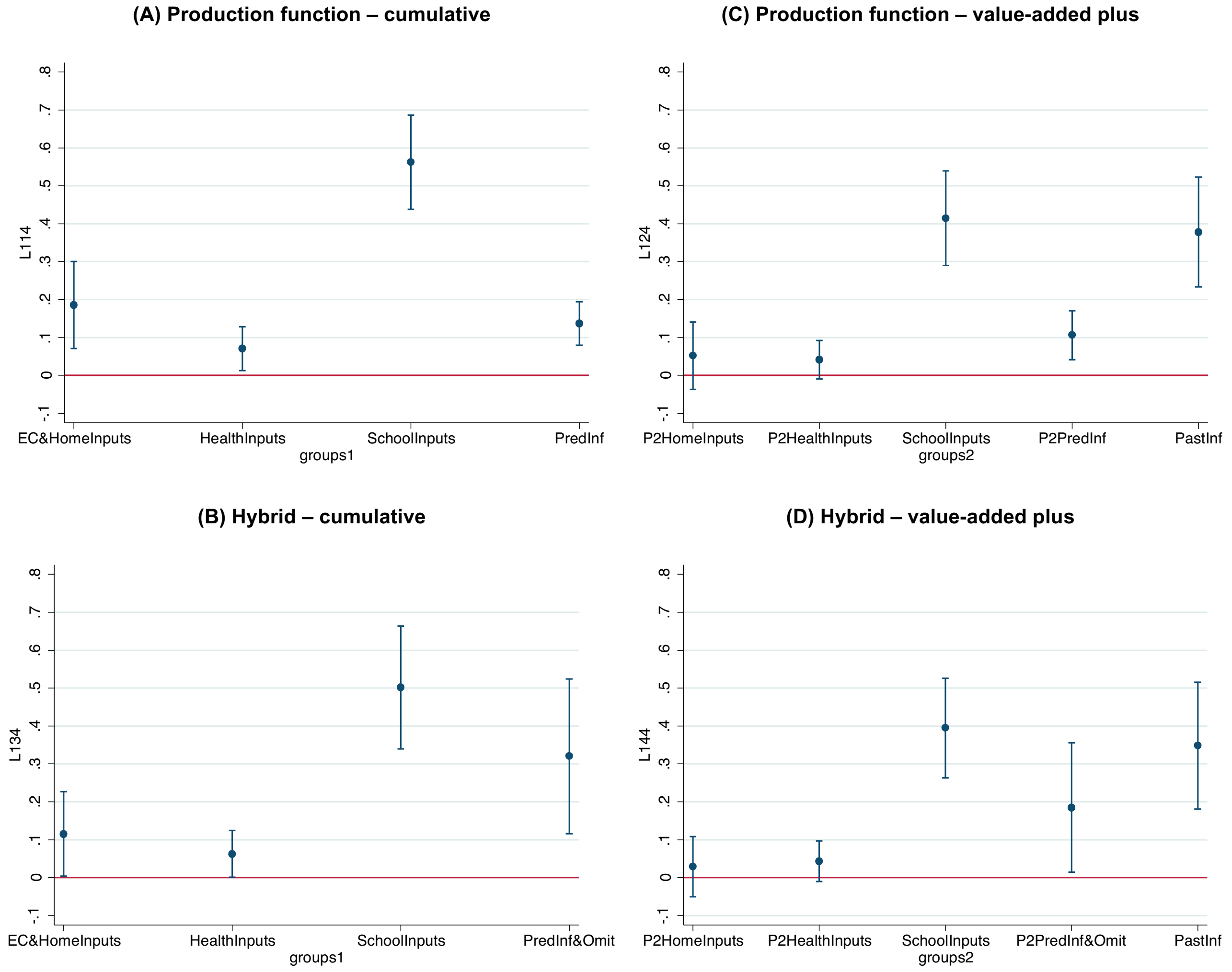
Table A4. *Coefficient estimates for the four specifications excluding school inputs from the school survey and working with the complete sample* continued

VARIABLES	(1) Prod.Fn-CU	(2) Prod.Fn-VAP	(3) Hybrid-CU	(4) Hybrid-VAP
Child lived in urban area (yes = 1)			0.453***	0.246***
			(0.085)	(0.083)
Average household total income			0.048**	0.027*
			(0.018)	(0.014)
Average household size			0.009	0.008
			(0.014)	(0.013)
Proportion of male siblings			-0.012	-0.029
			(0.069)	(0.052)
Child birth order			-0.076***	-0.051**
			(0.020)	(0.021)
Caregiver aspiration for child was university education (yes = 1)			0.216***	0.147***
			(0.034)	(0.033)
Standardised raw PPVT score (round 2)		0.447***		0.399***
		(0.027)		(0.030)
Constant	-1.848**	-0.135	-1.667*	-0.249
	(0.753)	(0.750)	(0.807)	(0.813)
Observations	1,561	1,561	1,561	1,561
R-squared	0.425	0.542	0.473	0.557

Robust standard errors in parentheses

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

Figure A1. *Normalised contributions to the urban–rural gap in cognitive skill at age 8 (point estimates and 95% confidence intervals) – including school inputs from the school survey and teacher pedagogical content knowledge*



(A) Production function – cumulative

(C) Production function – value-added plus

(B) Hybrid – cumulative

(D) Hybrid – value-added plus

Ho: $(\bar{z}_U - \bar{z}_R)'\psi = 0$; stat = 0.21, p-value = 0.053

Ho: $(\bar{z}_U - \bar{z}_R)'\tilde{\psi} = 0$; stat = 0.10, p-value = 0.350

Explaining the Urban–Rural Gap in Cognitive Achievement in Peru: The Role of Early Childhood Environments and School Influences

In Peru, students attending rural schools demonstrate extremely poor learning outcomes and obtain results significantly below those of students in urban schools. Because the process of cognitive skill formation is cumulative, differences in initial endowments, early environments and influences occurring later at home and at school can all play a role in shaping these gaps. This analysis aims at measuring the contribution of school and early childhood influences to the difference in cognitive development observed, at the age of 8, between urban and rural children in Peru. Previous decomposition exercises using Peruvian data on the indigenous–non-indigenous achievement gap, report results that favour the role of household characteristics over that of schools or community-level variables. This analysis contributes new evidence based on an unusually rich dataset and provided by a decomposition strategy less prone to biases than those used so far in the literature. Results indicate that between 35 and 40 per cent of the gap in cognitive skill between urban and rural 8-year-old children is related to differences in school inputs (years of schooling, school and teacher characteristics) received between the ages of 6 and 8. This contribution is similar to that of the learning and care environment to which the child was exposed up until the age of 5. The characteristics of rural schools have a direct connection with policy action because nearly all the supply of educational services in rural areas is public. Thus, efforts devoted to ensuring the characteristics of rural schools and teachers become more equal with those in urban areas should allow a significant reduction in the cognitive skill gap between urban and rural children by the time they reach Grade 3.

Young Lives
An International Study of Childhood Poverty

About Young Lives

Young Lives is an international study of childhood poverty, involving 12,000 children in 4 countries over 15 years. It is led by a team in the Department of International Development at the University of Oxford in association with research and policy partners in the 4 study countries: Ethiopia, India, Peru and Vietnam.

Through researching different aspects of children's lives, we seek to improve policies and programmes for children.

Young Lives Partners

Young Lives is coordinated by a small team based at the University of Oxford, led by Professor Jo Boyden.

- *Ethiopian Development Research Institute, Ethiopia*
- *Pankhurst Development Research and Consulting plc, Ethiopia*
- *Save the Children (Ethiopia programme)*
- *Centre for Economic and Social Studies, Hyderabad, India*
- *Save the Children India*
- *Sri Padmavathi Mahila Visvavidyalayam (Women's University), Andhra Pradesh, India*
- *Grupo de Análisis para el Desarollo (GRADE), Peru*
- *Instituto de Investigación Nutricional, Peru*
- *Centre for Analysis and Forecasting, Vietnamese Academy of Social Sciences, Vietnam*
- *General Statistics Office, Vietnam*
- *Oxford Department of International Development, University of Oxford, UK*

Contact:
Young Lives
Oxford Department of
International Development,
University of Oxford,
3 Mansfield Road,
Oxford OX1 3TB, UK
Tel: +44 (0)1865 281751
Email: younglives@younglives.org.uk
Website: www.younglives.org.uk

Young Lives
An International Study of Childhood Poverty